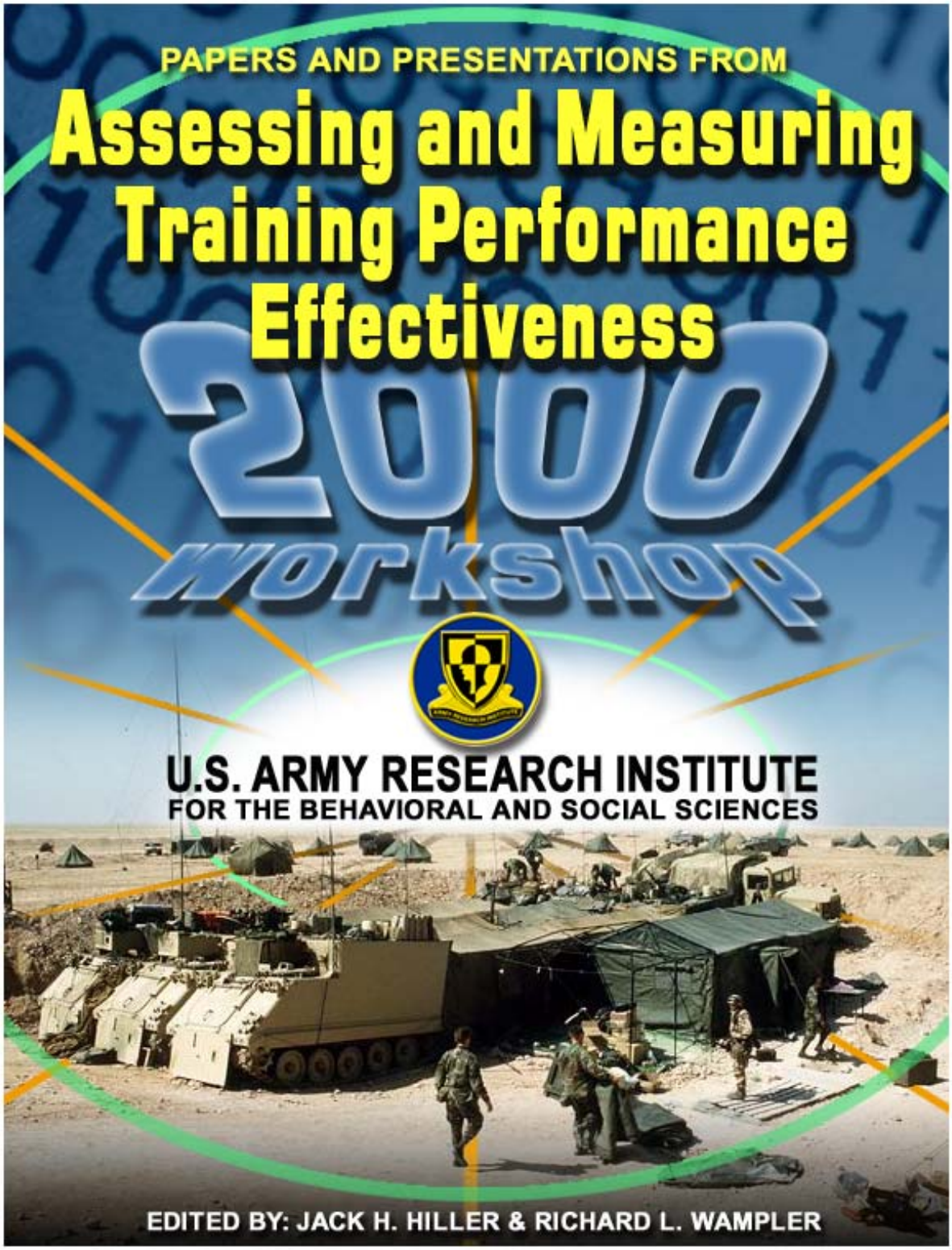

PAPERS AND PRESENTATIONS FROM
Assessing and Measuring Training Performance Effectiveness
2000 Workshop
U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES
EDITED BY: JACK H. HILLER & RICHARD L. WAMPLER

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
	Final	September 2000- January 2001

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Workshop on Assessing and Measuring Training Performance Effectiveness	DASW01-99-D-0013
	5b. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5c. PROJECT NUMBER
Jack H. Hiller and Richard L. Wampler (TRW)	5d. TASK NUMBER
	5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
TRW Inc. Systems & Information Technology Group One Federal Park Drive Faifax, VA 22033-4411 U.S. Army Research Institute for the Behavioral and Social Sciences ARI, Research and Advanced Concepts Office 5001 Eisenhower Avenue Alexandria, VA 22333	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333-5600	ARI
	11. MONITOR REPORT NUMBER Technical Report 1116

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Delivery Order Contracting Officer's Representative, Dr. Michael Drillings.

14. ABSTRACT *(Maximum 200 words)*:

This report provides documentation of the papers and briefings presented to workshop participants. The goals of the workshop were to provide key Army leaders with:

- A review of current state-of-the–art methods for training performance measurement
- Identification and clarification of measurement and assessment issues
- Recommended solutions and identification of essential research and development.

Eighty-five individuals from government agencies (both military and civilian), academia, and contractors attended the workshop held in Newport News, VA on 6 - 7 September 2000.

15. SUBJECT TERMS

Training Performance (human) Measurement Assessment Evaluation
Validation Reliability (measurement)

SECURITY CLASSIFICATION OF	19. LIMITATION OF	20. NUMBER	21. RESPONSIBLE PERSON

16. REPORT	17. ABSTRACT	18. THIS PAGE	ABSTRACT	OF PAGES	(Name and Telephone Number)
Unclassified	Unclassified	Unclassified	Unlimited		Dr. Michael Drillings (703) 617-8641

TABLE OF CONTENTS

Panel 5: Performance Measurement and Assessment Issues.

Introduction

Useful work in the physical and behavioral sciences, including engineering, is a product of the interaction of three distinct human activities:

1. INTUITION. An intuitive grasp of significant questions, issues, and topics provides a starting point. Intuitively recognizing key concepts, e.g., time, force, velocity, recognizing the means for observing and collecting data related to these concepts, and then collecting and analyzing such data distinguishes science and engineering, S&E, from pure philosophy and mathematics.

2. THEORY CONSTRUCTION TO ACHIEVE ALGORITHMIC COMPRESSIBILITY. The real trick in S&E is to be able to abstract from the virtually infinite data that may be collected to reflect the empirical world a manageable sample that will be useful for purposes of understanding phenomena, predicting future phenomena, inventing, and/or constructing formal models (e.g., theories or formulae).

3. MEASUREMENT. S&E, as explained above, make fundamental use of gathered data. However, not any data will do. Generally, visual observations supported by calibrated measuring devices (e.g., yardsticks, weighing scales, ohmmeters…) provide the most accurate, valid and reliable measures of phenomena to create useable data. By way of contrast, human measurement variables (e.g., perception of fragrances), often have weak measurement qualities, i.e., uncertain or low validity and reliability.

This workshop was concerned with measurement issues originating when training has been conducted, and human performance then needs to be measured to assess the effectiveness of the training.

Human performance is itself a complicated function of: the kinds of specific tasks to be performed, the wider environment or situation of performance, the conditions of performance including tools and equipment, the prior rest and physical conditioning of the trainees, the time interval between training completion and performance, intervening learning of potentially confusing tasks, and motivation to perform, as well as the goodness of the preparatory training. Furthermore, human performance is itself often difficult to observe, and objectively judge or rate.

Given the inherent complexity of measurement in training, and given that TRADOC faces especially difficult issues as the Army transitions to the future force, with it reliance on flexible leadership and multi-skilled soldiers, ARI and TRADOC decided that it would be helpful to hold a workshop to provide an overview of the state-of-the-art in training and instructional measurement, to hold focused panel discussions for selected key topic areas, and to develop recommendations for interventions or for essential R&D.

The workshop was held on the 6th and 7th of September, 2000 at Newport News, VA. Approximately 85 people attended. The Agenda is at Appendix A and the list of attendees is at

Appendix B. The papers prepared for this workshop follow in the body of this report. At Appendix C are the outbriefs from each of the five panels:

Panel 1. Proficiency Measurement in Technical Training Evaluation
Panel 2. Leadership Training and Education
Panel 3. Staff Training Assessment
Panel 4. Unit Collective Training
Panel 5. Performance Measurement and Assessment Issues

Briefings are reproduced in Appendix D.

It must be noted here that the idea for conducting this workshop came from Dr. Edgar M. Johnson, Director of ARI, who, along with Dr. Michael Drillings, Director of ARI's Research and Advanced Concepts Office, contributed greatly to the formation of the agenda and the summary recommendations which follow.

Summary Recommendations

The Army Transformation will benefit from systematic performance measurement and evaluation. Successful change requires measurement of trained performance and analyses of performance effectiveness to:

- Guide iterative or evolutionary development (i.e., design, test, revise, test),
- Ensure training and systems performance effectiveness,
- Validate effectiveness of training, Training Aids, Devices, Simulators, and Simulations (TADSS), and total systems performance.

Goals

Provide Army leaders with:

- A review of current state-of-the-art methods for individual, team, leader, staff, unit and soldier-machine systems performance measurement and evaluation,
- Identification and clarification of measurement issues,
- Recommended solutions or identification of essential R&D.

Recommendations

1. Army performance assessment should emphasize the use of convergent measures for the multiple dimensions of performance to be consistent with modern assessment practices. The additional strain on resources will be repaid with increased validity.

2. Operational units must validate the contents of distance learning (DL) courses to ensure their use and effectiveness. This action will have a high return on investment (ROI).

3. Training effectiveness must be validated by on-the-job performance measures. The results of the validation must be fed back to developers.

4. More frequent assessment of job task requirements and methods are needed to ensure that courses are brought into line with requirements created by new or revised doctrine and equipment.

5. Because of the need for soldiers to know how to operate specific variations of systems and specific versions of computer software, it has become important to assess accurately training accomplished and then to record what was learned and when in personnel and training databases.

6. GO-NO GO measurement scales may need to be expanded to enable measurement of mastery for more effective and efficient training. Research is also needed to examine trade-offs between mastery-level skill training (included optimal training methods) and the current practice of training many skills to only minimum levels of proficiency.

7. Routine assessment of leadership performance is essential for motivating leaders, generating feedback for self-development, focusing mentoring, informing personnel managers, and

evaluating the leader development system. ARI's leadership assessment toolkit will contribute to this capability, and help to overcome the existing suspicion of leader assessment practices.

8. The Army should provide feedback on leadership assessment during the after-action-review (AAR). A leadership AAR methodology for use by leaders and their assigned observer controllers (OCs) is now being explored by FORSCOM.

9. Mission Training Plans (MTPs) do not provide specific Measures of Performance (MOPs), and Measures of Effectiveness (MOEs). R&D is essential for improving this component of MTPs for analogue equipped units, for digital units, and for the Initial Brigade Combat Team (IBCT).

10. R&D is needed to create effective on-the-job training (OJT) performance measurement and recording tools. These tools will enable the development of better training plans and provide the basis for better personnel management.

Conclusion

Accurate measurement and assessment of performance, recorded, analyzed, and applied by training leaders create accountability and a source of information essential for assuring ROI. Evaluation needs to be highlighted as a cornerstone of sound training management.

Contributing Authors (Papers and Briefings)
(as they appear in this publication)

Jack H. Hiller, Ph.D., JD, is the Chief Scientist for the Information and Technical Services Division of TRW. He was formerly the Director of the Army's Personal Technologies Directorate (MANPRINT) on the Department of Army Staff, and Director of ARI's Training R&D Lab for nearly a decade. He was responsible for organizing the Workshop.

Scott E. Graham, Ph.D., is the Chief, ARI, Infantry Forces Research Unit, at Fort Benning, Georgia.

Michael D. Matthews, Ph.D., is formerly a research psychologist with ARI, Infantry Forces Research Unit, at Fort Benning, Georgia. He is currently a professor at the U.S. Military Academy at West Point, New York.

Robert A. Wisher, Ph.D., is a research psychologist with ARI in Alexandria, Virginia.

Henry Simpson, Ph.D., is a senior staff member in the Defense Manpower Data Center, in Monterey, California.

Stephen L. Goldberg, Ph.D., is the Chief, ARI, Simulator Systems Research Unit, in Orlando, Florida.

Andrew M. Rose, Ph.D., is the Chief Scientist at the American Institutes for Research in Washington DC.

Larry L. Meliza, Ph.D., is a research psychologist in ARI, Simulator Systems Research Unit, in Orlando, Florida.

Jerry Childs, Ph.D., is a research psychologist in TRW, Information & Technical Systems Division, in Albuquerque, New Mexico.

Millie Abell, Ph.D., is Chief of the Technologies Branch, Future Training Division, Assistant Deputy Chief of Staff – Training in Headquarters, TRADOC at Fort Monroe, Virginia.

Leonard (Don) Holder, LTG (R), is a senior program manager in TRW, Information & Technical Systems Division, in Killeen, Texas.

Chris Sargent, COL, is Director of the Center for Army Leadership at Ft Leavenworth, Kansas.

Michael Drillings, Ph.D., is Chief of the Research & Advanced Concepts Office in ARI, in Alexandria, Virginia.

William Mullen III, BG (R), is a senior program manager in TRW, Information & Technical Systems Division, in Monterey, California.

Marven Nickels, COL, is Director, Combined Arms and Services Staff School at Fort Leavenworth, Kansas.

Kathleen Quinkert, Ph.D., is Chief of the TRADOC – Scientific Coordination Office for ARI at Fort Monroe, Virginia.

John Johnston, COL (R), is a senior program manager in TRW, Information & Technical Systems Division, in Killeen, Texas.

Kent Ervin, COL, is Director of the Collective Training Directorate of the Combined Arms Center at Fort Leavenworth, Kansas.

Ward Keesling, Ph.D., is a research psychologist with PRC in Monterey, California.

Steve Ellison, MAJ, is an analyst in the TDA Division of the Assistant Deputy Chief of Staff – Training in Headquarters, TRADOC at Fort Monroe, Virginia.

Elizabeth Brady, Ph.D., is a research psychologist with ARI in Alexandria, Virginia.

Eva Baker, Ph.D., is Director of the UCLA National Center for Research on Evaluation, Standards, and Students Testing in Los Angeles, California.

Stephen J. Zaccaro is a professor in the Psychology Department of George Mason University in Fairfax, Virginia.

Lon E. (Bert) Maggart, MG (R), is Program Director, Advanced Learning Environments, Research Triangle Institute, Research Triangle Park, North Carolina.

Paul Rossmeissl, Ph.D., is a Senior Scientist at the American Institutes for Research in Washington DC.

Karol Ross, Ph.D., was formerly a research psychologist with the Human Research and Engineering Directorate, Army Research Lab at the Artillery Forces Research Unit, Fort Sill, Oklahoma.

Jim Lussier, Ph.D., is a research psychologist in ARI, Armored Forces Research Unit, at Fort Knox, Kentucky.

John Metzko, Ph.D., is a staff researcher in the Institute for Defense Analyses in Arlington, Virginia.

John Morrison, Ph.D., is a staff researcher in the Institute for Defense Analyses in Arlington, Virginia.

Frank J. Apicella is Technical Director of the Army Evaluation Center of the Army Test and Evaluation Command in Alexandria, Virginia.

Successfully Evaluating Training Devices in an Imperfect World

Jack H. Hiller, Ph.D., JD, Chief Scientist
[1] TRW, Information and Technology Services Division, S&ITG

ABSTRACT

In a recent article on methods for testing and evaluating the effectiveness of training devices and simulators, the authors (Boldovici and Kolasinski, 1997) identified Type II statistical error (i.e., failure to reject a false null hypothesis) to be a serious threat to the design of these tests and the interpretation of their results. Common problems identified were: low statistical power (e.g., when performance measures have "high" variability and the sample sizes are "small," tests for statistical significance will fail even though there are real population performance differences); failures to estimate statistical power; and failures to specify the testing results expected before data gathering that would support any decision to reject the null hypothesis of equality between experimental and control group performance (i.e., failure to specify the beta probability regions under the null hypothesis). Boldovici and Kolasinski explained that without having estimates of statistical power before conducting a test to evaluate the comparative performance of experimental and control training groups, sample sizes may be, "so small as to preclude finding differences between compared groups…" or so large that evaluation resources are "wasted." The purpose of this article is to explain an approach for designing a testing and evaluation strategy for new devices, especially complex collective training systems such as the Close Combat Tactical Trainer, when traditional methods of inferential statistics are not justified because of presumed low statistical power.

[1] The author thanks John Boldovici, Harold Wagner, Donald Headley, and F.J. Brown for their helpful comments on a draft. Dr. Henry Dubin contributed to a discussion reflected in the view of this paper, but is not responsible for its wording. Copies may be requested by EMAIL to Jack.Hiller@TRW.com

Successfully Evaluating Training Devices in an Imperfect World

Background

The article by Boldovici and Kolasinski focused on the Army's largest program for developing training simulation, the Close Combat Tactical Trainer (CCTT). The testing and evaluation for the CCTT will most likely be based on a single, major test of a mature design, because of the high costs for testing large Army units (i.e., armored companies) and the practical difficulties of scheduling the participation of these operational units. Since major training device tests are typically infrequent or one-shot affairs, evaluators will always lack the historical data that would be necessary for realistically estimating the probability of statistical power and Type II error. Thus, requiring device evaluators to estimate the probability of Type II error, as a prerequisite for evaluation, would logically prevent data collection or analysis of results. According to the logic of Boldovici and Kolasinski, evaluation would have to stop, or be viewed as wasting resources to generate unsupportable conclusions.

A contrasting methodological perspective may be found in Deming's chapter, "The Logic of Evaluation," in the *Handbook of Evaluation Research* (1975). Deming distinguishes between "enumerative " and "analytic" studies:

Effective use of statistical methods requires careful distinction between enumerative studies and analytic studies, with continual recognition of the limitations of statistical inference.[2] The aim of any study is to provide a basis for action. There are two broad types of action:

Enumerative—Action on the Frame.
Analytical—Action on the cause-system (process) that produced the frame and will produce more frames in the future. Page 57.

Although the use of Deming's distinction does not provide a perfect fit here, it applies well to distinguish between the major purpose of the approach proposed in this paper—Action on the Frame— and the Boldovici purpose, which applies to action on the cause. In other words, the approach taken in this paper emphasizes the value of validly conducted formative evaluation, in which the primary goal is to improve the training product (or to terminate work on it if the feedback from trainees, trainers and/or subject matter experts is too negative). In contrast, the Boldovici approach is summative in nature, seeking primarily to accept or reject a training product as an improvement over existing or alternative means of training support. Thus, this article will explain an approach for designing a formative testing and evaluation strategy, keyed to complex unit training systems, for new simulators and devices when traditional methods of

[2] Deming also cautions against the general application of null hypothesis testing for managing practical problems and denigrates resort to statistical power estimates as a guide for research: "We must face the fact that it is impossible to calculate from the data of an experiment the risk of making the wrong choice. The difficulty is that there is no statistical theory that will predict from data of the past what will happen under economic or physical conditions outside the range of the study. We can only be sure that conditions outside this range will be encountered. There is thus no such thing as the power of a statistical test. (These assertions conflict sharply with books and teaching on tests of hypotheses…)" Page 60. "The sad truth is that so-called tests of hypotheses, tutored well but not wisely in books and in teaching, are not helpful in practical problems, and as a system of logic, are misleading. " Page 62.

inferential statistics are not justified because of presumed low statistical power.

Sanity Checks

In 1968, as a new Ph.D. serving on active duty, I was assigned to the Army's Night Vision Lab. to assist with their recently adopted plan to include psychologists in the Lab's developmental and testing research programs for starlight and thermal imaging devices. A collection of charts was shown to me that related target viewing parameters (e.g., light or thermal target - background contrasts, light levels and thermal readings, target types and their distances from observers, etc.) to observer performances in judging detection and recognition of various targets. The data had been collected from a small number of soldiers (ten to twenty) acting as observers in a marsh at Warren Grove, New Jersey. The remarkable characteristic of the graphs was the smoothness of the curvilinear patterns. When I asked the responsible physicist how relatively small sets of data (five to twenty observations) could have generated such smooth curves, I was shown a French curve as his tool.

Duly outraged by this unscientific analytic procedure, I was led to consult with a leading expert in sensor technologies (Lucian Bibberman). When I solicited his support for rejecting the graphs and the research methodology that relied on small sample observations, he provided the following two arguments against my rejection. First, before these small data samples were collected, there were no field data whatsoever to provide empirical evidence of real world performance for these new and expensive imaging systems. Second, although the use of the French curve was itself unjustified, models for performance capabilities of the devices had been constructed from basic principles of physics, and the data served to validate the models, or show the need for refinements or rejection. He asserted that the data collected generally conformed to predictions from the models.

Thus, I was reminded of earlier lessons in research design. Some data may be better than none (only a qualified endorsement for data here, since bad data will mislead). And data collected to disconfirm a predictive model (the antithesis of "dust bowl" empiricism) have far greater utility than randomly collected data. In common parlance, people will typically suggest a "sanity" check when a new theory or predictive model has been proposed. Any procurement program costing a billion dollars, and possibly critical to the national defense, surely merits a sanity check, even though application of inferential statistics is not justified by statistical power estimates. Thus, I am proposing an approach to training device evaluation that checks for sanity, not inferential formal statistics (**SNIFS**).[3] The issue that needs to be managed is how to frame an evaluation strategy to create useful data and avoid any strategy that would produce measures of random noise, i.e., the unbounded ballpark estimates that Boldovici and Kolasinski properly criticized.

[3] This approach is suggested for application only when statistical power is an unresolveable issue. Unlike some others who would abandon significance testing (e.g., Schmidt 1996), I continue to believe that hypothesis testing with the aid of classical probability distributions or bootstrapped distributions (Lepage & Billard 1992) provides useful information for shaping theory.

Prototypical Evaluation Strategies for Producing Useless Ballpark Performance Estimates

This section proceeds by working through evaluation strategies that have actually been considered for application to the CCTT. The analysis of these strategies generally follows the classical treatment by Campbell and Stanley (1963).

SINGLE EXPERIMENTAL SAMPLE ELEMENT WITH A POSTTEST. One armor company would be trained on a test mission with the CCTT substituted entirely for training in the field. Where to start? Regarding statistical analysis, a single sample element fails to provide any basis for estimating the generalizability of performance that will be found across different companies. A single sample element is most likely drawn from the most dense region of a normal probability distribution, but there are no data here to check if the underlying distribution is normal; indeed, it might be a flat, equiprobability distribution that was sampled.[4] Regarding *external validity*, substitution of CCTT for all field training makes no sense -- no responsible field commander would manage his training program this way if CCTT were available, so the experimental use of CCTT lacks generalizability to real unit training.

Regarding *internal validity*, the lack of any control group (i.e., here an armor company training without the use of the CCTT) deprives the analyst of any ability to answer one of three fundamental evaluation questions: how does trained performance for the experimental group (with CCTT) compare to trained performance for the control group? (The other fundamental questions are: a) how well does the system meet its engineering objectives, e.g., picture and sound quality, mean time to failure, etc? and b) can trainees acquire or maintain the skills identified as enabling and terminal learning objectives?) The rationale offered to justify the one-shot group study has been that the experimental group will be evaluated against absolute Army training standards, a form of criterion referenced testing. This rationale is defeated by the unchecked assumption that conventionally trained units (an implicit comparison group) are routinely trained to standards in the same time frame.

SINGLE SAMPLE ELEMENT WITH PRETEST AND POSTTEST. To overcome the lack of a basis for comparison in the one-shot study, this design uses a pretest to measure the effects of the experimental treatment (i.e., use of CCTT). External validity is jeopardized by the effects of the pretest, since units do not ordinarily take a formal test before training, and never take any test that exactly resembles the posttest. Internal validity is also jeopardized by the similarity of the pretest to the posttest, since the pretest contributes directly to learning regardless of the use of the treatment (CCTT).

One virtue *claimed* for this design is its ability to produce a measure of gain associated with use of the experimental treatment (the CCTT). However, Cronbach and Furby (1970) have argued convincingly against use of gain scores. Since the typical question requires a comparison of treatment effects (i.e., which group achieves higher post-training scores, the control or experimental treatment group ?), the best analysis is a straight forward comparison of results when the comparison groups have been created through an effective random assignment

[4] However, as the sample size is expanded from one to many elements, the Central Limit Theorem would apply to assure normality of the sampling distribution for the sample mean values.

procedure. When the groups have not been formed by a valid random assignment procedure and do not, therefore, accurately represent their population, then no statistical techniques can be applied to correct for the misassignments, i.e., neither gain score analyses nor any other forms of covariance adjustments can correct the problem.

Multiple sample units should be included in the experimental treatment and in the control treatment to enable an examination of the shape of the performance distributions and examination of the variability of performance within the two or more treatment groups. Pretesting should be employed only where there are concerns about lack of comparability among sample units (i.e., individual armor companies here) assigned to the different treatment groups; a finding of significant and/or "large" differences in performance for the experimental and control groups on a pretest requires a new effort to randomly assign units to treatment groups – covariance procedures cannot correct for assignment of superior units to one group and inferior units to another.

How to Form a Ballpark within the Universe.

Early thinking on how to evaluate CCTT relied on a design strategy in which the control and experimental treatment groups would be formed by having multiple armor companies randomly assigned to each treatment, and that is statistically useful. However, according to one proposal, the control group was to be constrained to train only in the field, and the experimental group was constrained to train only in CCTT; furthermore, each individual sample unit (i.e., each armor company) was left entirely free to construct its own training program, given that it either stayed in the field or in its CCTT cabinets. Finally, the units with the CCTT simulators were totally free to use them for whatever task training they selected in any manner they chose. Now, all of this unconstrained variability is very bad for constructing a finite ballpark.

DAMAGE TO EXTERNAL VALIDITY. Modern armor units have a variety of training devices and simulators available. For example, the computer-based Unit Conduct of Fire Trainer (UCOFT) has been validated as a high fidelity trainer for tank gunnery. To have the control group restricted from using the UCOFT or other devices would be a distortion of normal unit training that would fail to generalize beyond such an artificial evaluation.

LARGE SOURCES OF UNCONSTRAINED TRAINING PROGRAM VARIABILITY. There are hundreds of individual and collective skills to be learned in any of the major missions for armor companies, e.g., Deliberate Attack. These many tasks vary considerably on trainability in the CCTT. For example, tank drivers can feel the influence of gravity and momentum in the real world, and cues from these forces act as feedback when driving, but of course the CCTT lacks this natural form of performance feedback. Driving skills are best learned through a combination of a specially designed high fidelity driving simulator and field training. Before a young driver is subjected to hours of intense training in the CCTT combat training simulator, where bad habits may be learned to mastery, there is a need for intense, high fidelity training. Skill training sequence and the amount of training employed, to achieve a given level of mastery, are critical training management features. Given the many tasks to be trained, the many ways to attempt training, variations in sequence and amount, unit training programs allow for nearly infinite variation.

The design of doctrinally approved unit training program models is a responsibility of the Army's Training and Doctrine Command (TRADOC) which it has fulfilled, with targeted funding from the Army's Director of Training and research assistance from the Army Research Institute (ARI).[5] Highly experienced military trainers, retired and currently serving, constructed detailed training program models that specified sequence, duration or repetitions, and expected quality of alternative training media, e.g., named simulators and devices and different field training conditions. To ensure the utility of the models, they were translated into training program guides which units can use for constructing their own training programs (Hiller, Wallace, Marcy, and Akam, 1995). The TRADOC worked as a co-developer to develop these training management guides (called Combined Arms Training Strategies, CATS), which may also be applied as a framework for conducting the CCTT evaluation. The ballpark is thus seen here to be forming from the CATS framework.

USE OF CCTT FOR HIGH PAYOFF STRUCTURED TRAINING. Given the variable quality of task training offered by CCTT (good daytime gunnery; excellent command and control day or night; low quality driving fidelity, etc.) there is an opportunity to enhance the payoff from CCTT training by capitalizing on its strengths and avoiding its weaknesses. The practice of carefully scripting training scenarios to train specified tasks (e.g., call for fire on a high priority threat), while avoiding execution of tasks in a low fidelity environment (e.g., movement on or near minefields where the warning cues are not realistically portrayed) has been termed structured training (Brown, 1994). The concept of structured training has been adopted by TRADOC, and has provided the basis for development of a library of structured training scenarios for use in CCTT training.
With the provision of a structured training program for using the CCTT and a formalized unit training management guide (CATS), the ballpark is now visible, and worthy of attracting evaluators.[6]

Varieties of Useful Measures of System Performance Effectiveness

Testing and evaluation is limited to the kinds of relevant data or measures that may be collected. In the case of CCTT, there is potentially a richness of data. Three dimensions of evaluative data are identified here: Results or Outcomes, Task Performance Process and Procedures, and Subject Matter Expert Evaluation.

TASK AND MISSION PERFORMANCE OUTCOMES. The first thought on how to evaluate a training system or program is to check if the trainees are mission capable, as demonstrated by their accomplishment of mission assignments or contributing tasks. For the Army, there are

[5] ARI organized and managed a contract program performed by the BDM Corporation as the primary contractor, with support from PRC Inc.

[6] The program for developing this library was funded through extraordinary support from the Army's Deputy Under Secretary for Operations Research, Walter Hollis, and the Commanding General of the Operational Test and Evaluation Command, MG Lehowicz, with the work directed by the Army Research Institute Research Unit at the Armor School, and with contract support from the Human Resources Research Organization and the BDM Corporation.

three bottom-line measures, which are: seizing terrain, holding terrain, and the ratio of enemy killed to friendly killed, the traditional casualty exchange ratio. These measures have an obvious relevance for evaluation, but in the context of CCTT will be limited by practical considerations to small sample sizes at the company echelon, and thus offer weak statistical power. Measurement reliability may be improved, however, by conducting repeated tests for the few units tested and then averaging test performances.[7] For example, in a study of the relationship between Ground OPTEMPO and unit performance, as measured by casualty exchange ratios, a sample of only 16 combined arms brigades was available. Each of the brigades was naturally divided into two combined arms task forces. Each task force fought four or five battles on defense at the National Training Center (NTC) producing a single casualty exchange ratio as the performance outcome measure for each battle. The four or five ratios were averaged for each task force, and then the two averages for the two task forces in each brigade were averaged. The 16 averaged casualty exchange ratios were then correlated with the Ground OPTEMPO expended by the 16 brigades in the six months preceding their visit to the NTC. The correlation for defensive missions was $r = .64$, $p < .01$, demonstrating the power of averaging, since any one battle appears to present considerable random variation (Hiller, McFann, and Lehowicz 1994).

PROCESS MEASUREMENT. Army doctrine specifies how tasks are to be performed to meet standards where a given procedure is believed to be optimal or where standardization contributes to training and performance efficiency. Observation and measurement of how tasks are performed is generally more informative and useful than mere outcomes, especially when outcomes are contingent on a number of uncontrolled or poorly controlled variables, such as enemy preparation and effectiveness, changes in weather, etc. (discussed in Hiller 1987 and 1994). While observation of performance in the field is often difficult to arrange, observation of performance within the computer-based CCTT may be relatively easy and precise. For example, a gunnery task that requires the tank commander rapidly to direct aimed fire on a visible threat may be hard to observe and record in the field, but may be done well within the CCTT.

Process measurement of complex unit collective tasks may unfortunately suffer from rater unreliability with no warnings to evaluators. The short "war story" that follows illustrates this problem. In the early 1980s, I led a team of training developers as we reintroduced small unit (infantry squad) battle drills to the Army's training literature (Hiller, Hardy, and Meliza 1982). The evaluation team members consisted of an outstanding Lieutenant Colonel and Platoon Sergeant of infantry (Jones and Jackson), an experimental psychologist (Meliza), and a civilian researcher with many years experience with the infantry (Hardy). Each of the raters, who shared in writing and carefully editing all of the drills, independently scored infantry squads as the squads performed after training to standard, by the squads' own reckoning. As it turned out, each of the raters produced a distinctive rating pattern that defied inter-rater reliability. LTC Jones scored almost all performances as NO GO. PSG Jackson scored almost all performances as GO. Meliza scored GO and NO GO equally often. Hardy felt that the terrain and enemy conditions described for effective training in the drills had not been met and refused to score performance as meaningless. Thus, even knowledgeable raters need to be calibrated, and the Army's Operational Test and Evaluation Command is working to develop training to control inter-rater reliability.

[7] It is useful to recognize that this concept serves only to increase the reliability of performance measurement for the few units included in a sample, but does not increase the generalizability of results.

SUBJECT MATTER EXPERT EVALUATION. There are three sources or kinds of subject matter experts (SMEs) that can provide valuable objective and subjective information for purposes of testing and evaluation. These are SMEs who serve as:

a) independent observers of training,
b) trainees with sufficient expertise to provide valid evaluative information, and
c) trainers.

Active and retired service members possessing a depth of real-world experience with the functions and tasks to be trained can be tasked to independently observe and evaluate training (including their non-disruptive questioning of trainees). These SMEs should be directed to identify any training features that are good, with appropriate explanations. SMEs should also be asked to indicate any features of a device that produces poor training or even negative transfer to live performance. Asking SMEs to comment on the value and quality of a device or simulator is traditional, but for a simulator as complex as the CCTT, evaluators should solicit evaluations according to explicitly identified system components or features. A short list would include the following.

Performance Cues. Clarity and fidelity of the terrain and objects presented on viewing screens for daylight and thermal displays. Audibility and fidelity of communications, and realism of noise.
Response Controls. Fidelity of the feel and responsiveness of operator controls.
CCTT Performance Feedback. Fidelity of system reaction to operator responses. For example, does the tank slide back when a vertical climb is too steep or stop when crashing a hillside?

Summary Performance Feedback System for Supporting After Action Reviews. Does the system cover the most important sources of information, provide sufficiently easy and rapid access, and present the feedback in a manner that is easy to comprehend and apply to learning and selection of follow-on training activities? See Meliza, Bessemer, and Hiller (1994) for a comprehensive description of an experimental system.

Training Management. Appropriateness and usability of the Combined Arms Training Strategy and the CCTT Lesson Library (e.g., do users find that the ground OPTEMPO saved, say 80 miles per year, is tolerable, and the large number of simulated miles, say an extra 1000 per year in CCTT, provides a training advantage). Usability of the CCTT training management system for tracking each unit's training history of demonstrated strengths and weaknesses (including tracking of unit leaders and members by name to meet special training needs created by personnel turbulence and turnover), and usability of the lesson selection and scheduling tools. The reader may have noted that simple descriptive summary statistics will be adequate for the SME evaluation. Thus a straightforward sanity check is sufficient for this evaluative dimension.

Analysis of Testing Data

Results from the three evaluative dimensions should be compared for consistency and sensibility. Outcome results from performance of missions may be explained by results from the procedural task performance evaluation, and results from the task performance evaluation

may be explained by the SME's evaluation. Corroboration of testing and evaluation results from such multiple sources (a form of convergent validation) will provide greater credibility for conclusions than the mere rejection of a single null hypothesis (Lykken, 1968). Furthermore, if the results found seem unexplainable and implausible, then the finding may not be made acceptable simply because the null hypothesis was rejected (see Lykken's discussion of the uselessness of significance testing to validate implausible hypotheses).

When descriptive data summaries for the three dimensions consistently support a conclusion that the experimental training program is equal or comparable to conventional training (which is the CCTT training program goal under the CATS), then the evaluation may fairly conclude that the experimental program has met its goal. If the experimental program's results are inferior to conventional training, then the first question to be answered concerns the magnitude and specific areas of necessary improvement. For small apparent differences between the experimental and conventional systems, no further analysis would be justified, but for apparently large differences, a test of the null hypothesis is desirable to avoid drawing conclusions from chance results. A failure to reject the null hypothesis should force attention to issues concerning Type II error, as discussed by Boldovici and Kolasinski, and/or a check for a possible problem with the sampling procedure used to form the experimental and control groups.

When the evaluation results consistently show deficits in the performance of members of the experimental training group, regardless of magnitude, then detailed examination of the specific deficits is warranted to identify faults to be corrected.

Given a conclusion that substantial or consistent deficits have been found, then analyses should be conducted to identify the source(s) in the experimental device, the training scenarios or techniques, and the training program's management. For example, consider a circumstance in which the conventional group scored an average of 88% tasks correctly performed and the experimental group scored 80%, with the null hypothesis for task performance differences between the conventional and experimental groups rejected at $p<.01$.[8] The difference for the averages at 10% may be regarded as substantial, but the training device would not be discarded. Instead, evaluators and training developers would search to find any tasks systematically under-performed in the experimental group to fix the device or the training program. After corrections have been made, testing would be resumed as is customary for iterative training development methods. The development of experimental devices will be terminated only when serious or fatal problems have been found that defy correction, or when the corrections cost too much.

Analysis of OPTEMPO-Simulator Tradeoffs

Budgetary constraints dictated that the CCTT's procurement and maintenance costs would be amortized by reducing Ground Operating Tempo (OPTEMPO, i.e., field training mileage). Given that CCTT offers high quality training, it might be possible to reduce field training below the amortization value to save money. There is, however, an intractable methodological problem confronting any evaluation of substitutability of simulator training for

[8] As a practical matter, task performance profiles would be examined, regardless of any failures to reject the null hypothesis, to find specific meaningful problems and affordable solutions.

field training when the evaluation would be conducted with trainees who enter the evaluation after having experienced field training.

The soldiers participating in any contemporaneous evaluation will have previously trained for years in the field, most particularly the leadership -- the Non Commissioned Officers, Captains, Majors, Lieutenant Colonels and full Colonels. The command and control training of these leaders in the CCTT simulators, and their learning from the simulation, will be conditioned by their years of field experience. They will tend to avoid making mistakes in the simulator that they remember from previous field training and perform according to long memories. Thus, potential inadequacies in the simulation will be "overlooked" by using the simulator as a memory prompt. Later field performance during testing can be expected to benefit from training in the simulator, but a major contribution to the simulator training will have been rekindled memories of earlier field training.

Consider now test results in which the experimental group, who used the CCTT with a substantial reduction in field training, performed as well or better than the field training control group. The budgeteers might interpret such results to mean that OPTEMPO may be substantially reduced. However, we can see that such a one shot evaluation can not be used to accurately estimate a harmless reduction in OPTEMPO. Accurate evaluation of the proper mix of field and simulator-based training can only be accomplished over an extended period of time.

Major New Applications of the Billion Dollar CCTT

We have until this point taken for granted the purpose of the CCTT, and that is a topic of neglect. Evaluation is fundamentally driven by the purpose of the object to be evaluated, and its major components. The underlying technology for the CCTT has been termed appropriately Distributed Interactive Simulation (DIS), and early in the development of DIS technology it was recognized as having a potential for doing more than training routine unit collective mission skills. Tasks that are especially **difficult** to train, **dangerous** to train, and **expensive** to train in the field could be mastered (e.g., multi-service and joint fire support of ground operations). Highly efficient focused training for **leader battle command skills** could be conducted by substituting CCTT in conjunction with war game models (e.g., JANUS) for training a portion of expensive large scale maneuvers. The CCTT is recognized to have a latent capability for enabling units to **rehearse** the execution of specific missions on objective terrain by incorporating up to date maps and satellite produced photographic imagery. Furthermore, the technology has the latent capability to enable **research** on execution of existing tactics, techniques and procedures by creating an archival data base of unit performance that could be researched for "Lessons Learned," as had been done earlier by the Army Research Institute for data collected from training at the National Training Center, Ft. Irwin, California. Finally, a powerful application of DIS technology would be its use in examining and **developing** newly conceptualized doctrine, communications, and weapons systems, **with the user in the loop** to realistically assess the feasibility and value of proposed innovations. These future applications of CCTT technology should be addressed by future evaluations.

Conclusions

Use of the Combined Arms Training Strategies (CATS) and use of the CCTT library of structured training scenarios in testing and evaluating the CCTT will radically constrain variability of results, as compared to the original, unconstrained evaluation concepts. With variability of unit train-up programs controlled by application of doctrinally approved unit training management models (i.e., CATS), and variability in the use of CCTT constrained by its doctrinally approved library, a plausible basis for conducting a sanity check (SNIFS) of CCTT data is established. Based on these considerations, evaluating CCTT data by SNIFS may present a practical solution for the problem created by low statistical power. Once in the ballpark, descriptive statistics may be collected for all three evaluative dimensions and used to form reasoned judgments on the value of complex, device-based simulations, such as provided by the CCTT. The application of SME judgment was given a prominent role here for evaluation and for identifying specific problem fixes and improvements.

Ultimately, there is no compelling need to determine by a test of the null hypothesis if the conventional and experimentally trained groups are different, for they surely are. Their differences are not at issue unless the experimental group is found to be "substantially" *inferior* by outcome measures ("interocular significance, a result that hits you between the eyes," Scriven 1997, page 20) or by observations of procedural task performance mistakes, or by SME judgments to reject the device for stated reasons. In all likelihood, any "substantial" unit performance deficits or serious dislikes found and judged to be practically significant would stimulate redesign efforts and re-testing --- and not determine any wholesale rejection of the experimental device. The salient issues for evaluators concern robust effects or differences and analysis of their sources to fix problems or capitalize on successes.

Once the experimental group performance is found to equal or exceed the conventional group, the training device would be judged effective, and the training management strategy (CATS) and its lesson library would be considered validated, so that only the financial costs of the new device and its new training capabilities (as described above for major new applications) would constitute the proper grounds for procurement decisions. Thus, hypothesis testing has at most only marginal relevance after a substantial investment has been made in a new training (or operational) technology, and problems may be corrected by affordable solutions. Furthermore, following Lykken's (1968) reasoning on the value of significance testing for experimentation, we may conclude:

> the finding of statistical significance is perhaps the least important attribute of a good [evaluation]; it is *never* a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an [evaluation] report ought to be published. (Page 158).

References

Boldovici, J. A. & Kolasinski, E. M. (1997). How to make decisions about the effectiveness of device-based training: elaborations on what everybody knows. *Military Psychology, 9, 121-135.*

Brown, F.J. (1994). A new training paradigm. In Holz, R.F., Hiller, J.H., and McFann, H.M. (Eds.). *Determinants of effective unit performance: research on measuring and managing unit training readiness* (pp. 281-298*).* Alexandria VA: U.S. Army Research Institute.

Campbell, D. T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In Gage, N.L. (ed.) *Handbook of Research on Teaching* (pp. 171-246). Washington: D. C.: American Educational Research Association.

Cronbach, L. J. & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74, 68-80.*

Deming, W.E., (1975) The logic of evaluation. In Struening, E.L. and Guttentag, M (Eds.), *Handbook of Evaluation Research* (pp. 53-68). Sage Publications, Beverly Hills

Hiller, J.H. (1987). Deriving useful lessons from combat simulations. *Defense Management Journal,* Second and Third Quarter, 28-33 (terminal issue).

Hiller, J.H. (1994). Assessment of training readiness. In Holz, R.F., Hiller, J.H., and McFann, H.M. (Eds.), *Determinants of effective unit performance: research on measuring and managing unit training readiness* (pp. 299-306). Alexandria VA: U.S. Army Research Institute.

Hiller, J.H., Hardy, G. D., & Meliza, L.L. (1982). *Guideline for designing Drill Training Packages.* ARI Research Product 84-11.

Hiller, J.H., McFann, H.M., and Lehowicz, L.G. (1994). Does OPTEMPO increase unit readiness? In Holz, R.F., Hiller, J.H., and McFann, H.M. (Eds.), *Determinants of effective unit performance: research on measuring and managing unit training readiness* (pp. 71-79*).* Alexandria VA: U.S. Army Research Institute.

Hiller, J.H., Wallace, S. W., Marcy, S. C., & Akam, R. B. (1995). Development of a Force XXI training management strategy. *Army Research, Development, and Acquisition,* May-June, 10-12.

Lepage, R. & Billard, L., eds.(1992). *Exploring the limits of bootstrap.* New York: John Wiley and Sons.

Lykken, D. T. (1968). Statistical Significance in Psychological Research. *Psychological Bulletin,* 70, 151-159.

Meliza, L.L., Bessemer, D.W., & Hiller, J.H. (1994). Providing unit training feedback in the distributed interactive simulation environment. In Holz, R.F., Hiller, J.H., and McFann, H.M. (Eds.), *Determinants of effective unit performance: research on measuring and managing unit training readiness* (pp. 257-280). Alexandria VA: U.S. Army Research Institute.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Bulletin,* I, 115-

129.

Scriven, M. (1997). "The Vision Thing": Educational Research and AERA in the 21st Century, Part 1: Competing Visions of What Educational Researchers Should Do. *Educational Researcher*, 26, page 18.

Modeling and Measuring Situation Awareness

Scott E. Graham
U.S. Army Research Institute
Infantry Forces Research Unit
Fort Benning, GA

Michael D. Matthews
United States Military Academy
Dept of Behavioral Sciences & Leadership
West Point, NY

At times it seems that "Enhanced Situation Awareness (SA)" has become the rallying cry for today's combat developers. "Enhanced SA" promises to make fighting wars, and related ventures, faster, cleaner, …more efficient. We hear that the "seamless integration" of disparate digital technologies will produce a "common picture of the battlefield," which will give soldiers and leaders "perfect situation awareness." That all sounds appealing. Knowing the location, and the intent, of all friendlies, enemies, and civilians will be of great help; clearly casualties and fratricides should be reduced. Whether these digital dreams will be realized is yet to be determined. In the mean time, you might want to hold onto your compass -- just to be safe.

Most of the focus on SA has been on the design of digital architectures, digital displays, and a smorgasbord of sensors. By contrast, we in the training, leader development, and soldier (TLS) business see one of the most critical SA requirements as being how to develop leaders who can exploit the new digital information. That is, how do we train leaders and soldiers to use the digital equipment and the plethora of information to make better decisions. In part, we know this is going to require valid SA measurement approaches that can assess SA processes, outcomes, and related decisions. This paper discusses work we have done to address some of the unresolved TLS SA issues, including an Infantry-focused SA model, new SA measurement techniques, and on-going SA research.

One of the problems in dealing with SA is that there is no commonly agreed upon meaning for the term. The Army tends to refer to SA as knowing where you and your buddies are, where the enemy is, and the location of civilians. This definition is useful, but lacks the breadth and precision needed for theoretical models. Pew (1998) notes that SA definitions tend to be circular and vacuous, i.e., better performance implies better SA and vise versa. The hardware/technical crowd often talks of SA in terms of bandwidth, or as being a particular device, e.g., "Here is our SA display." We know that SA involves much more than that. Maggart and Hubal (1999), for example, say that SA enables a commander to (1) place current battlefield events into context, (2) readily share a portrayal of the situation with staff and subordinates, and (3) predict, expect, and prepare for future states and actions. In short, SA is a set of related cognitive and perceptual processes, not a digital system or device. A good SA model should incorporate the full scope of SA processes and outcomes.

The U.S. Army Research Institute (ARI) has been working in the area of SA for several years, beginning with an Infantry Situation Awareness Workshop that we hosted at Fort Benning, GA in September 1998. The workshop objectives were to: (1) develop SA requirements and performance measures for Infantry combatants and teams; (2) establish a dialogue between cognitive and behavioral researchers and Infantry warfighters; and (3) identify requirements for future training, leader development, and soldier research. The papers from the

workshop were published as an ARI book (Graham & Matthews, 1999) and can be found on the web at www.ari.army.mil.

While the majority of the ARI SA research is being conducted under the Infantry Forces Research Unit's (IFRU) Training Modernization workpackage, we have received some additional funds from the Director of Bio Systems, Defense Research and Engineering (DDR&E) to follow up on the Infantry SA Conference. Much of the work being discussed in this paper, including the Infantry-focused Situation Awareness Model was developed as part of the DDR&E sponsored project (Endsley, Holder, Leibrecht, Garland, Wampler, and Matthews, 2000). The ARI, TRW, SA Technologies, and now, the U.S. Military Academy team is continuing to work together in the SA area.

Individual SA Model

Much of the SA research and the corresponding SA models have focused on fixed wing aircraft pilot issues. Our goal has been to develop an SA model that combines the dynamics of the Army/Infantry environment with sound theoretical perspectives of SA and human behavior. Our model centers around Endsley's (1988) definition of SA, "Situation Awareness is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." Endsley characterizes SA as having three levels: (1) Perception, (2) Comprehension, and (3) Projection of the future. This conceptualization is similar to that recently forwarded by the Commanding General, 4th Infantry Division (ID), the Army's first digital division, that differentiates situational awareness, situational understanding, and situational dominance. Following the 4th ID scheme, situational awareness corresponds to perception or the lowest level of processing. In this paper, and for the most experts in the field, SA is a term that refers to whole continuum of cognitive processing -- from perception to the highest cognitive processes.

We wanted to make sure our model was capable of handling the full dynamics of the Infantry battlefield environment. Also, while the focus of the model was originally on Infantry, most of the factors identified readily apply to all ground forces. In contrast to the relatively simple environment of an aircraft cockpit, the Infantry environment is much more complex, involving multiple transport platforms and all types of terrain, including urban terrain. Infantry units also are by definition groups of individuals, with differing abilities and dispositions. Informing, coordinating, and commanding Infantry units across a dispersed battlefield involves different SA-related processes than those required to command an aircraft or ships. Infantry forces are also unique in their close contact with the civilian population. Soldiers must perceive and interpret subtle cues in a foreign culture. Furthermore, increased SA capability will tend to produce greater unit dispersion and movement rates, the result of which can be increased danger, stress, and fatigue. For excellent insights into the complexity of the modern Infantry battlefield and SA-related issues, we strongly suggest reading *Black Hawk Down,* by Mark Bowden (1999).

We thought it essential that the Infantry-focused SA model be consistent with sound, practical wisdom regarding warfighter dynamics and SA. For example, we know SA to be significantly affected by one's experience, to include individual and unit training experiences.

We also know SA is affected by psychological factors, such as trust and cohesion. Furthermore, despite claims that future SA systems can provide perfect situation awareness, we know that uncertainty can never be totally dispelled. We know that SA skills can be developed. Army training should expose leaders to increasing amounts and complexity of information, and should stress their capacity to identify and understand key SA elements in a wide variety of tactical situations. We also believe that making "Quality of SA" a standard feature of after action reviews (AAR) should result in high payoff.

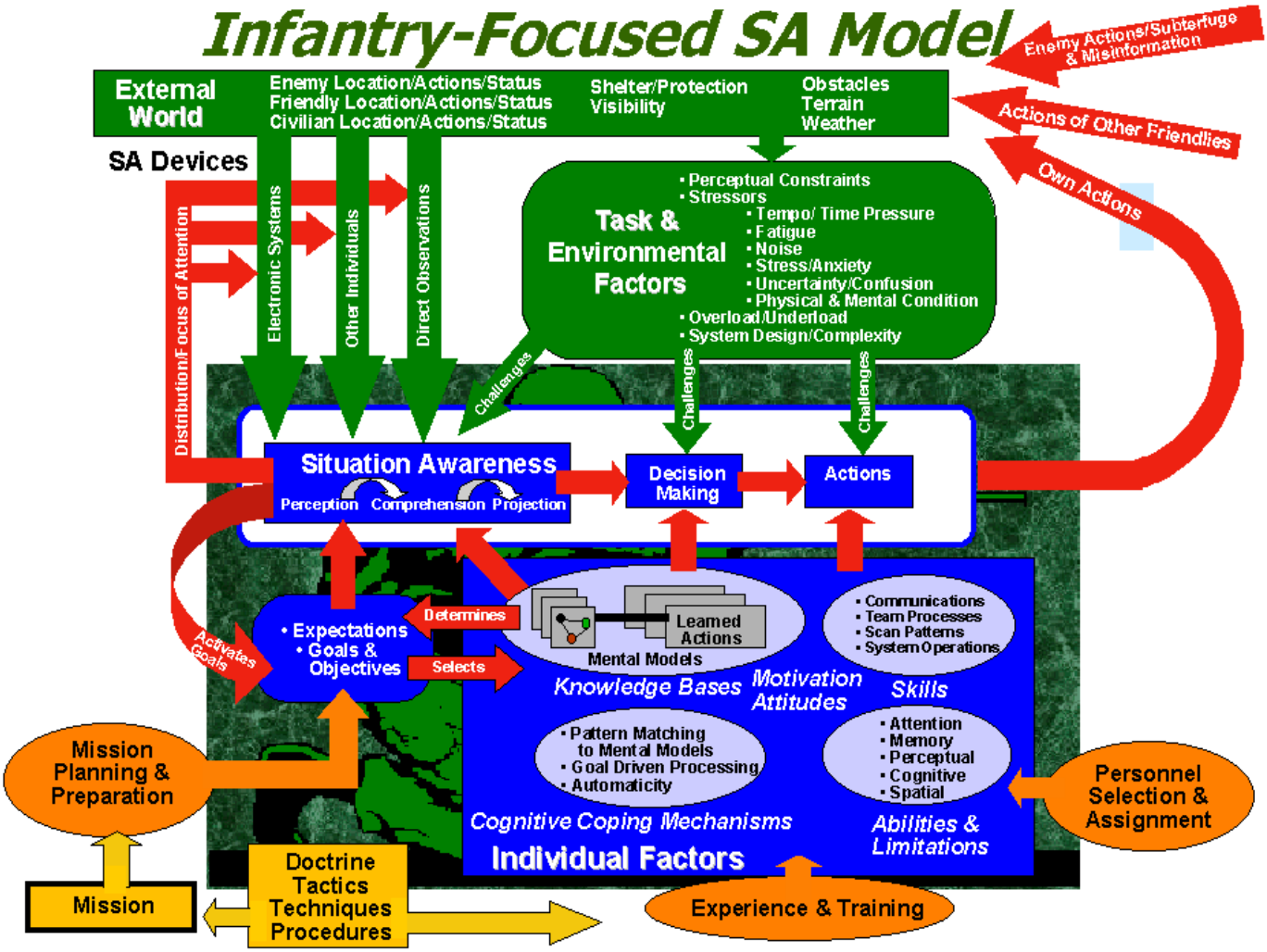


Figure 1. Infantry-focused model of individual situation awareness.

Figure 1 shows the Infantry-focused model for individual SA (from Endsley et. al. 2000). At the center of the model are the three levels of SA processing. The first level involves the perception of the status, attributes, and dynamics of relevant information in the environment. This will include information on the status of the enemy, friendlies, and civilians, as well as terrain features, obstacles, and the weather. The second level, comprehension, involves understanding the significance of the information in the context of the soldier's goals. For example, the importance of seeing a piece of terrain that has been disturbed may be understood

differently by experienced and inexperienced soldiers. The third and highest level of SA is the ability to project the future. Commanders with high levels of SA are able to project where and when the enemy will strike, how much time until they receive reinforcements or until the next artillery volley. Being able to project future conditions allows leaders to make quality decisions as to favorable courses of action.

Individuals derive SA from various sources, beginning with their direct observation of the external world, e.g., sight, hearing, smell, and tactile/kinesthetic senses. A second major source of information is that from communications with others. This includes both verbal communications, either direct or over radios, and non-verbal communications, for example, facial expressions or hand and arm signals. Various electronic sensors and displays will increasingly become information sources for SA, e.g., from global positioning system (GPS), night vision goggles, or Land Warrior displays. While the addition of each of these systems increases the likelihood that critical battlefield information is available to the soldier, there is a downside. All of these information sources compete for the soldier's limited attention and processing capability. Increased attention directed toward one source may result in less attention directed toward another source. For example, a soldier who is engrossed in analyzing information from a display may miss other important information around him.

Many factors can limit an individual's SA. The factors include perceptual constraints such as obstacles, noise and smoke. There can also be a lack of understanding of commander's intent, general confusion, and the enemy can deliberately conceal critical information or provide misinformation. Several types of stressors may also affect SA, including physical stressors, such as heat/cold, boredom, or fatigue, and social/psychological stressors, e.g., fear, anxiety or time pressure. Stressors can affect SA in various ways, including attentional narrowing, reduced information intake, and reduced working memory capacity. Research has also shown that both high and low workload can have a negative effect on SA.

Each soldier possesses certain abilities, skills, and knowledge bases that largely determine the quality of his or her SA. There is evidence, both empirical and anecdotal, that suggests some leaders are better at maintaining high SA than are others. These skills and abilities may be partially inherent, but they can also be enhanced by experience and training. The model identifies various cognitive processes involved in developing and maintaining good SA. As for the individual cognitive factors, included in the model, SA is going to be most restricted by limitations in attention and working memory. The model also identifies leverage points where the Army should focus its efforts for enhancing SA. These include the development of relevant knowledge bases for pattern matching, goal-directed processing, and automaticity of actions.

Shared SA

"Shared" or "team" SA is also extremely important, particularly for Army units. Shared SA requires shared mental models and shared goals, e.g., a clear and common understanding of the commander's intent. A battalion commander has shared SA requirements with his subordinate company commanders. While there will be high overlap in their SA requirements, a company commander's requirements will often be too detailed and situation specific for the battalion commander. Conversely, the battalion commander may be aware of "big picture"

issues that are generally beyond the purview of a company commander. Knowing where to draw the line, i.e., what to report and what to omit, is critical for successful SA in Army units.

The omission of critical information, either up or down the chain of command, can lead to catastrophic SA failures. Too much information can strain limited communication channels and thereby inhibit the communication of truly relevant information. The problem can be minimized to the extent that each person in the organization clearly understands the SA requirements of the others. This level of understanding only comes from considerable experience with other members of the team. A pervasive research issue is how to develop greater leader and team experience in less time.

Shared mental models can greatly facilitate communication and coordination in team settings. Team members with similar knowledge bases and cognitive mechanisms are more likely to interpret information the same way, as well as to make accurate projections about each other's decisions and actions. Without shared mental models, coordination and communication will likely take more time and effort, and will result in more lapses. Shared mental models can be enhanced by: (1) shared training, e.g., joint training or cross training on different job functions; (2) shared experiences, e.g., working together as a team or having similar experiences either together or individually; and (3) direct communications between team members to build up a shared mental model in advance of operations.

A number of studies, e.g., Klein, Zsambok, and Thordsen (1993), have examined factors affecting team processes that are related to shared SA. Some of the differences between effective and ineffective team processes include:

Ineffective Teams	Effective Teams
• SA black hole - One member misleads others	• Self-checking - Check against others at each step
• Don't share pertinent information - Group norm	• Coordinating - Get information from each other
• Failure to prioritize - Members go in own directions - Lose track of main goal	• Prioritizing - Set up contingencies (shared mental model)
• Over reliance on expectations - Unprepared to deal with false expectations	• Questioning - As a group

SA Measurement

Measuring SA in a combat environment poses significant challenges. Despite the importance of SA, it is nevertheless an inferred construct that does not directly translate to easily observable behaviors. Furthermore, SA is always going to be relative to "ground truth," and, at any moment, it may be difficult to know the actual conditions in a fluid combat environment. As shown in our model, there are a myriad of factors that affect SA, including information

complexity, rapidly changing information, information overload/underload, tempo, fatigue, noise, and stress. To complicate the measurement process, soldiers and leaders often rely on very subtle cues from the environment and other combatants.

Much of the interest in SA measurement surrounds the development of new digital information systems. In particular, there is the question as to whether these new systems actually enhance SA and to what degree. The systems, e.g., the Army Battle Command System, Land Warrior, or new video links, typically produce huge amounts of data. The problem becomes not the absence of information, but finding the appropriate information when it is needed. One problem is that it is sometimes difficult to determine what information the soldier or leader is attending to at any given time to produce the level of SA he or she may have.

There are a number of reasons why it is important for the Army to be able to measure SA. They include:

- Enhancing SA in Military Operations
  - What are the critical skills/abilities that lead to high SA?
  - What factors hinder SA the most?
  - How do soldiers maintain SA under harsh operational conditions?
  - What strategies lead to high SA?
  - How does SA develop within and between teams?

- Evaluation of system designs
  - Do new technologies actually improve SA?
  - Which aspects of SA are hurt by technology?

- Evaluation of training programs
  - How effective are new SA training techniques?

Figure 2 presents a model for organizing the various types of SA measurement techniques. The measurement approaches include both inferred and direct measures that can be applied across the SA continuum from perception through decision-making and action. A full description of each type of measure, with advantages, disadvantages, and application considerations, is included in Endsley, et. al., (2000). Direct objective measurement techniques, which query the individual for knowledge and understanding, have been used most extensively. This technique sometimes introduces probes during on-going exercises, but the more common approach is to freeze the exercises. During the freeze, the soldiers are asked detailed questions about the state of the environment. This method has been formalized by Endsley (1995) as the Situation Awareness Global Assessment Technique (SAGAT).

The use of complementary SA measurement techniques often yields the most complete and useful picture. Consider, for example, the assessment of the effects of a new global positioning system (GPS) on SA. Video recordings or eye tracking could be used to determine how much time the soldier time spent looking at the device, and whether the soldier used the GPS while being stationary or on the move. Direct measures of SA, such as SAGAT, could be used to ask for relevant information, e.g., current location, correct azimuth to next point, or the

location of the best tactical position. You could also measure soldier performance, such as time and accuracy of a decision, the ability to recover from system failures, or the speed and adherence to a prescribed route.

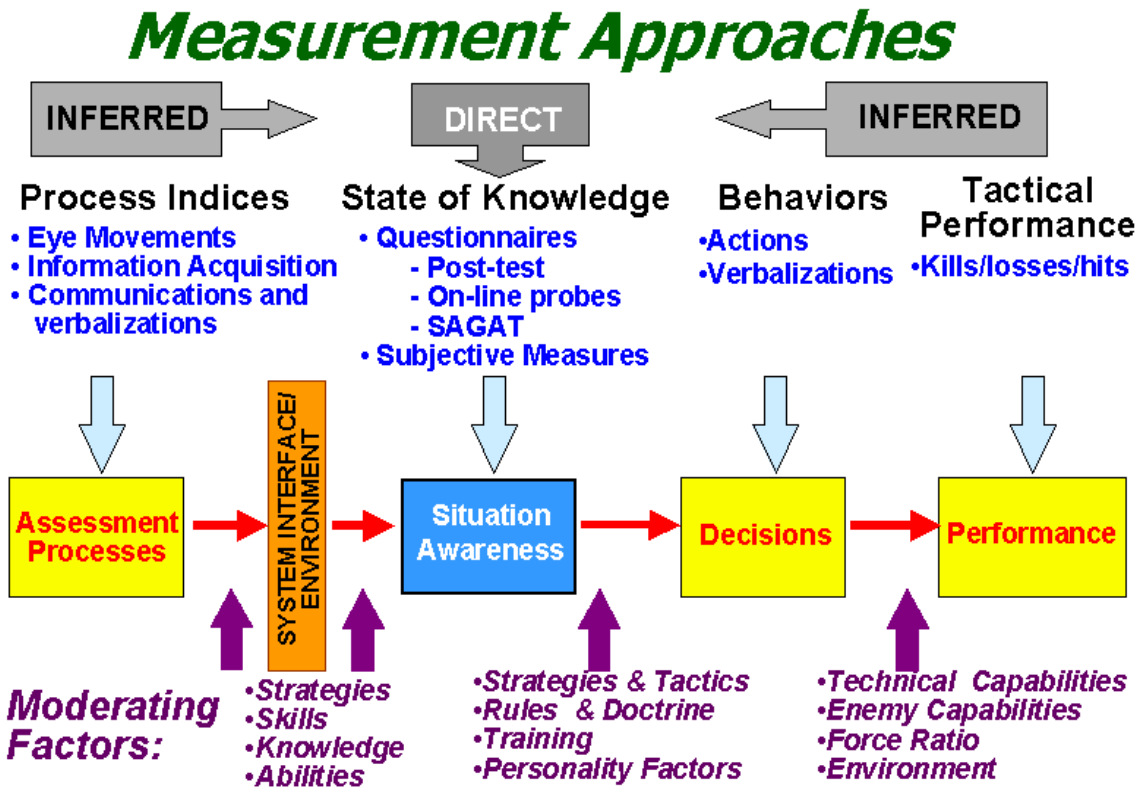


Figure 2. Process model of situation awareness measures.

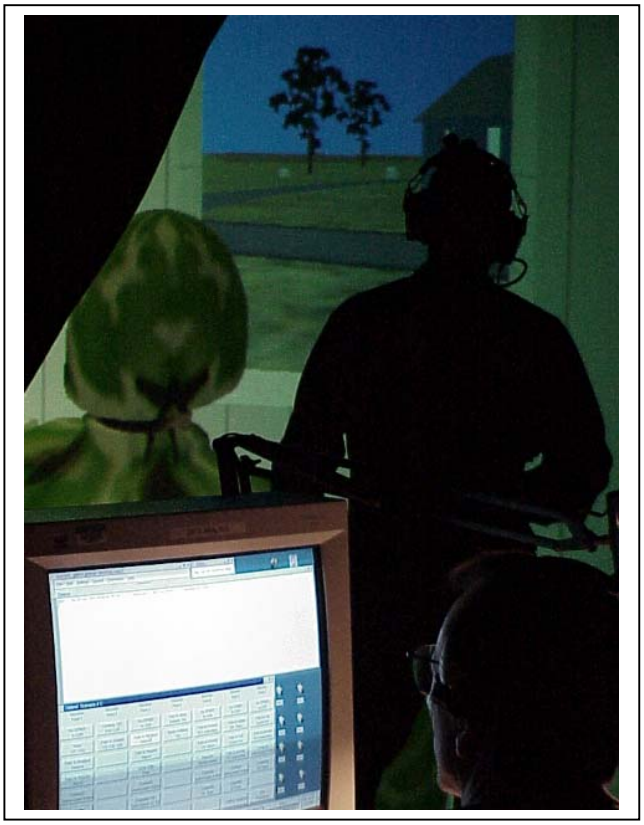
SA Measurement and Decision-Making Training

In a Virtual Environment

ARI, in partnership with the U.S. Army Simulation, Training, and Instrumentation Command (STRICOM), and the Army Research Laboratory (ARL) is working to develop effective methods for training small Infantry units in virtual environments (VE). We recently conducted an experiment whereby we trained Infantry platoon leader decision-making skills in the Squad Synthetic Environment (SSE). The SSE, a set of full-immersion simulators, is described in Pleban, Eakin, and Salter (2000).

Figure 3 shows a soldier being trained in the SSE during the July 2000 experiment. The objectives of the experiment were to (1) assess the capability of the SSE as a decision skills trainer, and (2) to develop and validate platoon leader SA measures.

SA Measures Development

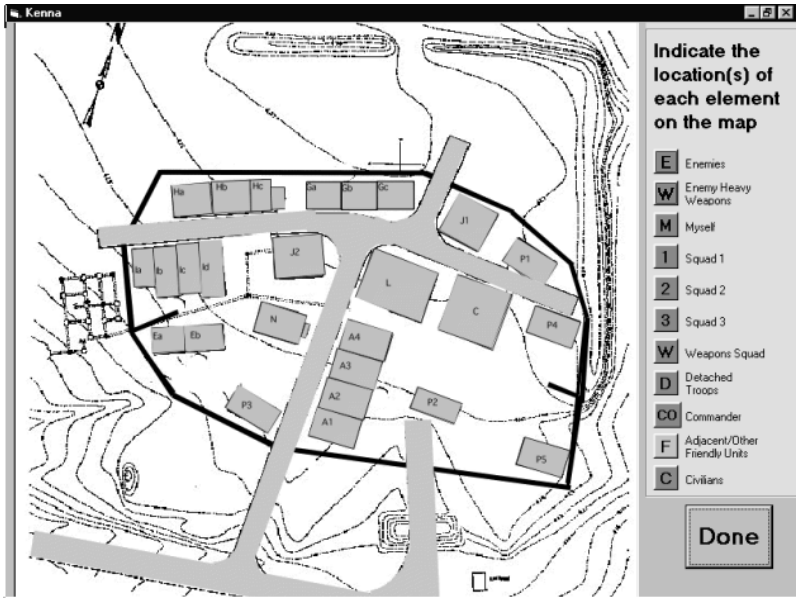


We began by conducting an SA requirements analysis for Infantry operations in urban terrain, as described in Matthews, Pleban, Endsley, and Strater (2000). The requirements analysis revealed seven key goals for attack and defend MOUT missions. These were: **avoid casualties, negate the enemy threat, movement (reach point X by time Y), assault through an objective, hold an objective, provide stability and support operations (SASO),** ***and*** **function in a team environment.** ***The seven goals were, in turn, further broken down into subgoals.***

Based on the SA requirements analysis, three different SA measurement devices were developed - a SAGAT-based measure, a situation awareness behaviorally anchored rating scale (SABARS), and a participant subjective SA questionnaire (PSAQ)

SAGAT. Twenty-one probe questions were developed, including questions about troop locations, available assets, and projection of the future. As previously discussed, the SAGAT procedure uses a freeze frame technique.

Four times during the selected scenarios the virtual simulation exercise was halted. A laptop computer was then rolled into the VE chamber on which platoon leader answered a series of randomized SAGAT questions. Figure 4 shows one of the SAGAT computer questions. The platoon leader's task was to drag the unit symbols to the appropriate positions on the Fort Benning McKenna MOUT site map.



SABARS. Expert observers rated the platoon leaders on 28 observable behaviors related to SA. Specifically, the SABARS

were five-point scales (with an additional response for "not applicable") on which the performance on specified behaviors was rated from "very poor," to "borderline," to "very good." Representative items included: "Solicits information from squad leaders," "Asks for pertinent intelligence information," "Uses assets to effectively assess environment," and "Projects future possibilities and creates contingency plans."

PSAQ. At the end of each scenario, the platoon leaders were asked to rate their own SA on a five-point scale. An example was, "Please circle the number that best describes how aware of the evolving situation you were during the scenario." Response options ranged from "Not aware of the situation" to "Completely aware of the situation. The platoon leaders were also given space to make open-ended comments about their SA.

Experimental Method and Analysis

Fourteen platoon leaders, seven experienced and seven inexperienced, were each given the opportunity to plan and execute four platoon level missions in the virtual urban environment. The platoon leaders were run individually, one platoon leader per day. The company commander, first sergeant, and squad leaders were confederates who followed scripted scenarios. The remaining platoon members, adjacent platoons, and the opposing force were computer-generated forces.

The first mission (Stability and Support Operation/Civil Disturbance) was used as a pre-test while the fourth mission (Secure Village/React to Downed Helicopter) was used as a post-test. During the middle two missions (Company Assault, Defend Town), the platoon leaders were coached on their decision-making and were given the SA assessment instruments. The missions ended with an after action review. Each of the scenarios contained four to six pre-determined decision points. In the "Company Assault," for example, the platoon leader had to make decisions about a failed breach attempt, a squad leader reporting that one of his squad members refused to fight, breach holes that were too high, and leaking containers in the midst of dead civilians.

The analyses will compare the decision-making performance with the objective and SA subjective measures. The report will be available in Nov 00 (Pleban, Endsley, Salter, Eakin, Strater, and Mattthews, in preparation). Interviews with the platoon leaders at the end of each day regarding the effectiveness of the virtual decision-making training were quite encouraging. Twelve of 14 platoon leaders said they thought the training improved their decision-making skills. Some representative comments were:

> "I learned more about decision making in my day here than in all of IOBC [Infantry Officer Basic Course]."
>
> "I was challenged by actual insertion in the virtual simulation vice "observing" JANUS. I was required to perform."
>
> "It gives leaders the opportunity to learn without jerking soldiers around. By the time leaders step in front of soldiers they will have some experience."

"Seeing the results of decisions I made greatly illustrates the effects/chaos of poor decisions or no decisions at all."

Other SA Research

We are starting to see some real progress in the modeling and measuring of SA for ground forces. In addition to the work described here, we have also been working with the ARL–Human Research and Engineering Directorate (HRED) and the Natick Soldier Center on measuring SA in the MOUT ACTD. There is also other SA work going on in the Smart Sensor Web program. In addition we currently have two small business innovative research (SBIR) projects that are about to begin that should provide useful insights and products. They are an OSD SBIR "Enhancing Situation Awareness in Military Operations," and an Army SBIR, "Assessing Decision-Making Skills in Virtual Environments." We in the TLS research community understand that research on ways to better train and measure SA is necessary if the Army and DoD is going to reap the full value of the new, and expensive digital systems. To succeed, we must continue to cooperate with one another in our research efforts and in communicating the value of this work to sponsors and stakeholders.

References

Bowden, M. (1999). *Black Hawk down: A story of modern war*. Philadelphia, PA: Atlantic Monthly.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37*(1), 65-84.

Endsley, M. R.., Holder, L. D., Leibrecht, B. C., Garland, D. J., Wampler, R. L., & Matthews, M. D. (2000). *Modeling and measuring situation awareness in the infantry operational environment.* (Research Report #1753). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Graham, S. E., & Matthews, M. D. (Eds.) (1999). *Infantry Situation Awareness: Papers from the 1998 Infantry Situation Awareness Workshop.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Klein, G. A., Zsambok, C. E., & Thordsen, M. L. (1993, April). Team decision training: Five myths and a model. *Military Review,* 36-42.

Maggart, L. E., & Hubal, R. (1999). A situation awareness model. In S. E. Graham & M. D. Matthews (Eds.), *Infantry situation awareness: Papers from the 1998 Infantry situation awareness workshop* (pp. 19-28). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Matthews, M. D., Pleban, R. J., Endsley, M. R., & Strater, L. (2000). *Measuring situation awareness in a virtual MOUT environment*. Paper presented at the Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium Conference, Savannah, GA.

Pew (1998). The State of Situation Awareness Measurement: Circa 1996. In Garland D. and Endsley M. (Eds). Experimental Analysis and Measurement of Situation Awareness, Mahwah, NJ.: Lawrence Erlbaum Associates, Inc.

Pleban, R. J., Eakin D. E. , & Salter, M. S. (2000). *Analysis of mission-based scenarios for training soldiers and small unit leaders in virtual environments*. (Research Report #1754). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Pleban, R. J., Endsley, M. R., Salter, M. S., Eakin, D. E., Strater, L, & Matthews, M. D. (In preparation). *Assessing Decision-Making and Situational Awareness Skills of Small Unit Leaders in Virtual Environments.* (ARI Research Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Measuring Performance In Distance Learning Environments

Robert A. Wisher
U.S. Army Research Institute

Introduction

Military training is concerned with increasing the capacity to perform military functions and tasks. For training specialized skills in the military, learning outcomes are established by doctrine and the criteria for minimally acceptable performance are generally set. Regardless of the delivery medium used or the instructional strategies employed, the learning outcome from training must ultimately translate to favorable performance.

The Training and Doctrine Command (TRADOC) is embarking on a major change in the delivery of individual and self-development training. By applying multiple media and networked delivery technologies, training is to move from classroom-centric instruction to a learner-centric model. With around-the-clock access to distance learning environments, soldiers will take on greater responsibility for learning facts, procedures, and complex skills as well as enhancing their teamwork skills. This change alters the manner in which future training will essentially be distributed and the methods by which training performance can optimally be measured.

TRADOC is transforming courses and configuring classrooms to accommodate the distributed training concept. As described in The Army Distance Learning Plan, over 525 courses are slated for redesign to a distance learning format by 2010. Related to this transformation, the National Guard Bureau established the Distributive Training Technology Project, which provides high-speed network links to armories in all states and territories (Bond & Pugh, 2000). The Army Reserve maintains a Distance Learning Futures Group which is examining alternatives to the traditional model of classroom training. Altogether, over 750 distance training facilities are planned throughout the Army, which would cover 95% of the total force, active and reserve components. In addition to these planned facilities, training will also be delivered to the workplace, to soldiers' residences, and to other sites apart from the traditional classroom.

At the same time, the Department of Defense (DoD) has established the Advanced Distributed Learning (ADL) initiative. This initiative grew out of a strategy to "harness the power of learning and information technologies to modernize education and training" (DUSD (R), 1999). ADL reflects the vision of ensuring "that DoD personnel have access to the highest quality education and training that can be tailored to their needs and delivered cost effectively, anytime and anywhere" (DUSD (R), 1999). The ADL initiative also marks a shift from classroom delivery to a model of training on demand through distributed learning technology. The advantages are increased accessibility to training, a reduction in long-term costs, the ability to change content rapidly, and a hoped for improvement in the overall product of training – performance.

An underlying assumption of these Army and DoD initiatives is that the quality of training shall be maintained whenever and wherever it is delivered to the service member.

Learning outcomes from a distance learning program must be on par with those from classroom instruction, if not better. The advantage of training from a distance, however, brings up the issue of measuring performance at a distance. How well can soldiers learn through distance learning technologies? What are the special requirements for measuring performance? Are there limiting factors? Do training policies need to be updated?

This paper examines these issues. It begins with a brief overview of distance learning and then discusses current paradigms for assessing learning outcomes. Several empirical examples from military applications of distance learning are reviewed. Shortcomings in current measurement practices are identified. Finally, the application of distance learning technologies as a measurement resource are presented along with considerations for future applications. These considerations included the development of performance metrics in the Shareable Courseware Object Reference Model being promulgated by the ADL initiative.

Distance Learning Overview

One definition of distance learning (DL), articulated by the Los Alamos National Laboratory, declares it to be structured learning that takes place without the physical presence of the instructor. This definition has been accepted by the U.S. Distance Learning Association and by military, government, education, and private sector activities concerned with the development and use of DL. Several defining characteristic of distance learning are: the physical separation of instructors and learners while instruction occurs, the presence of noncontiguous communication between student and teacher (through electronic media or print), and the volitional control of learning by the student rather than the instructor (Sherry, 1996). Table 1 provides a summary of the media being employed to deliver structured learning to the distant student. Forms of print, audio, and video represent early versions of distance learning. Computer-mediated conferencing and intelligent tutoring systems represent more recent advances. The Internet transcends all five categories.

Table 1. *Summary of Delivery Methods of Distance Learning*

PRINT	Delivered through mail, facsimile, or downloaded from the Internet		
	Correspondence study	*Training Manuals*	*Study Guides*
AUDIO	Delivered over cassette players, personal computer, telephone, radio, or the Internet		
	Audio cassettes	*Compact disc*	*Voice mail*
	Audio conferencing	*Radio broadcast*	
	Audio teletraining	*Streaming audio*	
VIDEO	Delivered over videocassette players, personal computer, satellite, microwave, fiber optic, cable, telephone, or the Internet		
	One-way video, 2-way audio	*CD-ROM*	*Streaming video*
	Two-way video, 2-way audio	*DVD*	*Videocassette*

Table 1. *Summary of Delivery Methods of Distance Learning (Continued)*

COMPUTER-MEDIATED CONFERENCING – Delivered through computer networks

Application sharing	*Bulletin board*	*E-mail*
Audiographics	*Chat Room*	*White Board*

COMPUTER-BASED TRAINING – Stand-alone (non-networked) training applications; audio and video as above.

Intelligent tutoring systems	***Embedded training***	***Electronic page turners***

A distance learning course applies one or some combination of these delivery methods. Instead of meeting at a centralized training location, DL offers instruction to students individually or in small groups situated at remote sites. The instruction can be synchronous in some applications or asynchronous in others. For synchronous delivery, instruction is projected from an origination site to two or more remote sites. As the breadth and reach of distance learning increases, combinations of delivery media will become more common and the Internet, or intranets, will assume a more central role in delivery and performance assessment.

Applications and Evaluations

According to the National Center for Education Statistics, there were 54,470 distance learning (DL) courses offered by institutions of higher education in the United States in 1998. An estimated 1,230 degree programs and 340 certificate programs were offered exclusively at a distance during 1997-98. The market for distance learning (which is also termed distributed learning, "DL" refers to either in this paper) is even larger when including training in government and industry. Each of the armed services and many large agencies have DL programs, which have been expanding in recent years.

The evaluation literature on DL has shied away from program effectiveness and focuses instead on usability, equipment quality, learner preferences, and learner satisfaction. Koble and Bunker (1997), for example, examined publication trends in a leading DL journal and found only 21% concerned evaluation of effectiveness. This is in part due to the time and cost necessary to perform a sound evaluation, particularly when students at remote sites must be factored into the sample. It also reflects a general lack of interest in measuring learning outcomes and performance. Many evaluations are conducted as an afterthought.

When evaluations are conducted, they are often done poorly. In a study concerning educational environments, Phipps and Merisotis (1999) point out that most research on distance learning does not control for extraneous variables nor use random assignment of subjects, and the validity and reliability of the instruments used to measure outcomes and attitudes are often questionable. In a parallel report on the literature as it pertains to training, Wisher and Champagne (2000) concluded: most research is anecdotal; when effectiveness is examined it is usually based on an ambiguous experimental design; when effectiveness is measured comparative results are only reported approximately one-third of the time; and when data are

reported there are analytic problems and errors in reporting that are often overlooked by researchers. In another examination of the literature, Joy and Garcia (2000) randomly selected representative samples of media comparison studies, and illustrated inadequacies of methodologies and conclusions.

Threats to Internal Validity in DL Research

The term "threats" has been used in research to represent alternative explanations for the results that are reported. A design that eliminates these threats is said to have high *internal validity*. That is, a DL program possesses internal validity if it can be established that the cause or treatment (i.e., use of DL media with a particular instructional strategy) was responsible for the effect or outcomes of the program (e.g., satisfaction, learning, performance). If internal validity cannot be demonstrated due to poor design, then the evaluator cannot conclude that the program "worked" (i.e., caused the higher performance).

Some researchers use designs that fail to eliminate many alternative reasons for the consequences of the training program. This usually occurs because of the lack of a comparison group and/or failure to obtain more than one measure of performance. Designs which use equivalent comparison groups or include pretest and posttest measures can make the results of DL studies more meaningful. Below are just a few of the threats to internal validity or alternative explanations for the results that can occur in studies of the training effectiveness of DL. In each case, the researcher may mistakenly attribute success or failure to the DL technology when it may have been due to another cause:

history - Changes in performance or attitude may be due to another specific event, other than the treatment or use of DL. For example, students may have learned the material from a source outside of class or were inspired to seek out other information outside of class.

maturation - Changes in performance or attitude measures may be due to students becoming less interested in the program or more fatigued over time.

mortality - Students with less ability, motivation, or time resources may become discouraged and drop out during the program so that the average posttest knowledge-based scores are higher than the average pretest scores.

test sensitization - Pretest measures may sensitize students to the knowledge-based items and they may score higher on the posttest regardless of the content of the training program.

Learners bring various degrees of prior knowledge to the learning process. Tobias (1994) determined that prior knowledge accounts for between 30 and 60 percent of explained variance in posttest scores. This knowledge is not always assessed prior to an instructional treatment, leading to a potential confounding in the interpretation of learning outcome data. Disappointingly, 50% of the evaluations relating DL to a training outcome use variations of a posttest-only design, whereby students were given a test of knowledge following the administration of the DL-based course (Wisher & Champagne, 2000). This design is generally uninterpretable due to the lack of a pretest measure of knowledge. Furthermore, half of the posttest-only designs did not use a comparison group, students who were not administered the course via DL. Threats to the internal validity of studies reported in the DL evaluation literature abound, limiting any overarching conclusions that can be drawn about its impact on performance.

Performance and Learning Outcomes

Factors influencing individual performance have been studied in the laboratory, in educational settings, and in the workplace. There have been problems in generalizing findings from one environment to another (Fleishman and Quaintance, 1984). For example, research in the experimental laboratory is difficult to interpret for use in workplace settings due to the absence of common task dimensions. The focus of the current review will be on performance in educational or workplace settings rather than laboratory environments. In this paper, the interest is on examining the effects of different learning conditions on task performance, specifically distance learning conditions compared to conventional classroom conditions.

Task performance may be measured immediately after training, such as through a hands-on test or a written knowledge test, or on the job, through an assessment of performance on specific tasks. Directly observed performance assessments, of course, can make a stronger case as to whether or not DL influences performance. Written tests, however, have a correlational coefficient of r = .62 with hands-on testing, as evidenced during the Army's Project A effort (Campbell, Campbell, Rumsey & Edwards, 1985). This means that written tests account for only 38% of the explained variance in hands-on performance tests. When examining the affects of DL on performance, results based on written knowledge tests should be contemplated with this correlation in mind.

A fundamental question is whether performance should be defined in terms of behavior or results of behavior (Smith, 1976). In the Army, the quality of task performance is key to understanding the capabilities for job performance. Hence, performance should be judged in terms of behaviors related to carry out military tasks and functions. Job performance in the Army, however, should not be equated to task performance. Studies have demonstrated, for example, that enlisted infantrymen spend less than half of their time performing the technical tasks for which they have been trained (Bialek, Zapf & McGuire, 1977). In determining the effectiveness of a learning condition, performance judgments must be based on the tasks that were trained under conditions that resemble the demands of the workplace.

Evaluations in Education and Industry

In educational settings, early evaluations of distance learning were mostly descriptive case studies that focused on learner satisfaction (OTA, 1989). They were often conducted as an afterthought and relied on reaction questionnaires that were often unreliable or not representative of the students involved. The focus was on student perceptions or immediate educational outcomes. Linking either of these variables to performance-oriented measures was largely ignored. For researchers interested in understanding the relationship between DL and subsequent performance, little could be gained.

In a meta-analysis of training outcomes from 34 studies, Alliger, Tannenbaum, Bennet, Traver, and Shotland (1997) found no evidence of a relationship between affective reactions by

learners to training and learning outcome measures. Utility judgments by learners fared better, but accounted for a small percentage of the variability of outcome measures. In a report by the National Research Council, evidence suggested that peoples' assessment of what they know or remember in laboratory studies can be seriously flawed, particularly when using one indicator, such as recognition, to predict another, such as performance (Druckman & Bjork, 1994).

Evaluations of training in workplace settings would be expected to link outcome variables to measures of job performance or productivity. However, many of the published evaluations of DL in business and industry are only summative in nature. Perhaps this is due to a reluctance by businesses to inform competitors on the details of applying DL and its contribution to their bottom lines. Examples of such summative reports include online training at Sprint, through the intranet-based Sprint University of Excellence (Harsha, 2000), e-learning practices for United Airlines' for training 10,000 customer service agents (Kiser, 2000), and satellite-based training broadcasts of Home Depot Television for familiarizing employees on new product information and customer service practices (Sims, 2000). More than 75% of Dell Computer Corporation's internal training is offered online via the company's intranet, but performance details are not available. From these summary reports and many like them, little can be gained from the business literature to deepen our understanding of how DL affects performance. The military, in contrast, has been more forthcoming with the details of evaluating distance learning, although the majority of reports (over 80% according to Walsh, Gibson, Miller, & Hsieh, (1996)) reported video teletraining as the DL technology rather than the new genre of Web-based learning tools entering the marketplace.

Evaluation Framework

Linking the outcomes in DL to subsequent performance on a task or on the job requires an organizing framework. A popular framework for evaluating training outcomes is the generic Kirkpatrick model (Kirkpatrick, 1984). The model has four levels of evaluation, three of which correspond to performance. It has been broadly applied in the literature, and it is relevant to DL. The model will be summarized here and then several examples of its application in military DL programs will be described.

Level I - Reaction Measures. Reaction measures refer to an individual's perception of some aspect of a training program, such as the quality of the video, the effectiveness of the instructor, or the overall quality of the program. These are largely affective reactions to the particulars of a course. Reaction measures are very common in the research literature but they demonstrate little correspondence to performance.

Level II - Learning Measures. Learning measures offer a more objective assessment of the knowledge and skills acquired during a training program. Knowledge refers to the facts, principles, rules, and procedures that were taught. It is generally measured through paper-and-pencil tests. Skills generally refer to the application, or transfer, of what was acquired in the classroom to a time and event dependent environment such as the workplace. Skills are generally measured through hands-on performance tests or situational exercises. The measurement of performance during or immediately upon completion of training is an example of a learning measure.

Level III - Behavioral Measures. Behavioral criteria are concerned with the follow-up performance of the participant in another environment, such as a workplace setting. The issue is whether what was learned in the training transferred to the workplace. For example, consider a case where certain soldiers participated in a weeklong DL course related to training digital skills for operating a battlefield system. The event may have received favorable reactions from the soldiers and may have increased their immediate knowledge and skill as indicated by a learning measure. However, if there is no improvement in later job performance as measured through, say, supervisory ratings, then the DL training cannot be declared a complete success. Behavioral measures require a period of time, weeks or months, before the effectiveness of training can be judged, and it is helpful to include a comparison group. The resources to implement a behavioral measure, especially the time factor, can be high.

Level IV - Results Measures. Results criteria are similar to behavioral criteria in that they are also concerned with the performance, but at an organizational level. A classic example is a sales training event. As in the previous example, a sales training event may have had favorable reactions from the participants, and measures of their learning were positive. But if there was no comparative gain in sales within the region, then on the basis of a results criterion the training event was unsuccessful. This measure also requires data collection over an extended period.

Related to these four levels are other measures of interest to the Army, such as the long-term affect on a career or the return on investment from training costs. The acknowledgement of successful performance in a Level II or Level III evaluation (or quasi evaluations) could boost an individual's confidence and motivation, leading to a more productive career progression and a longer career, a benefit which might go undetected in the constricted temporal window of an evaluation program. For return on investment, one must also keep in mind the depreciating value that training can exhibit due not only to skill decay (Wisher, Sabol & Ellis, 1999) but due also to the obsolescence of specific skills over time (Gordon, 2000). This latter point is particularly relevant to the rapidly evolving digital skill domains.

Learner Satisfaction

The most common measure used in the DL literature is, unfortunately, the reaction measure. For example, Walsh et al., (1996) concluded that for evaluations of DL in training environments, objective learning measures were used in only 36% of the cases. Learner satisfaction with courses is a common use of a reaction measure (Level I). It offers an interesting juxtaposition with the findings from learner achievement. The research suggests that learner satisfaction and achievement are independent (Payne, 1999). The degree to which a student is satisfied or unsatisfied with a DL course does not affect his or her level of achievement in the course. Similarly, a student's level of achievement does not influence his or her satisfaction with a particular course. This finding has been quite consistent. The primary evidence for this finding derives from the research literature on interactive video teletraining as reviewed by Payne (1999).

One military study that reflects this finding is that of Simpson, Wetzel and Pugh (1993). In this study, learner attitudes and training effectiveness for live instruction and six forms of video teletraining

were measured for over 700 students. The results indicated no significant differences in learner attitude or learning outcome between instructional formats. Based on reviews of numerous studies and meta-analysis, Payne (1999) concluded that "learner attitudes do not appear to impact learner achievement…. learner achievement does not appear to impact learner attitudes" (p. 11).

Distance Learning in Military Environments

Studies of distance learning in the Army have demonstrated positive results, if one believes that learning outcomes equivalent to those of a classroom represent a "positive result." This general finding should not be surprising since most early trials of DL were videoteletraining (VTT) implemented as a copycat form of the traditional classroom. For example, in the Florida Videoteletraining Project with soldiers (n=99) from the reserve component, two-way interactive video was applied for training three military occupational specialties (unit administrative specialist, unit supply specialist, and basic military police). Dependent variables were standard, criterion-based proficiency and achievement tests. The end-of training scores demonstrated learning outcomes equivalent to soldiers trained in a resident mode (Bramble and Martin, 1995). Other media for delivering distance learning in the military have also demonstrated the equivalent-performance effect.

Computer-Mediated Conferencing

The application of the DL delivery medium of computer mediated conferencing was applied to the Engineering Officer Advance Course. Fourteen reservists served as the DL group and the comparison group was constituted from final exam scores (n=339) at the resident site as well as a subset of resident students (n=49) for purposes of assessing demographics and perceptions at the resident site. The results showed no difference between resident and distance learning students on objective learning measures (Phelps, Ashworth & Hahn, 1991). The distance learning course was projected to cost less than the resident version when conducted over ten iterations.

Audioteletraining

In a study that measured the cost effectiveness of the DL delivery medium of audioteletraining, favorable results of equal effectiveness at a lower cost were reported for the training of unit clerks during a three week course (Wisher & Priest, 1998). Here, performance was measured through hands-on exercises in which soldiers performed a clerical task, such as completing a Survivor Benefit Form, and then faxed the results to the Army National Guard (ARNG) Professional Education Center in Arkansas. Instructors received the fax, graded the results, and returned a Go/No-Go score to the soldier. Of the 16 such performance measures obtained in this manner, the audioteletraining group (n=118) had a first time Go rate of 94% compared to a first time go rate of 86% for the comparison group (n=107). This difference was statistically significant ($t=4.7$, $p < .001$). The overall Go rate for both groups after multiple tries was 100%, thus the result of overall "equivalent performance." Based on the yearly training load, however, the audioteletraining version of the course demonstrated an annual cost avoidance of $300,000 due to savings in travel and per diem.

Audiographics

In a study using audiographics as a DL delivery medium, Wisher and Curnow (1999) reported on an application for a four-day course on computer security (i.e., cyber attacks, computer emergency response teams, etc.) conducted by the Army Land Information Warfare Activity. Audiographics refers to a medium in which a visual image is accompanied by the instructor's voice. Unlike audioteletraining, audiographics allows the instructors to present, annotate and manipulate the visual image. In this research, computer graphic images were displayed to remote sites using a T.120 data conferencing standard. Two-way audio communication in synchrony with the images supported the delivery. Thus, students at remote sites viewed a PowerPoint slide presentation, controlled by the instructor, while listening to the instructor's lecture over the audio bridge.

The audiographics originated from the Army Reserve Readiness and Training Center at Fort McCoy. Seven remote sites (n=107) participated in the DL version of the course. A comparison group (n=108) received the training in a traditional classroom at Fort Belvoir. The results on an objective written examination demonstrated scores of 88% for the DL group and 87% for the classroom group, no significant difference in course performance.

Other Services

Findings from the other services also reflect the effect of equivalent performance. Several relevant studies were conducted by Doug Wetzel and colleagues at the Navy Personnel Research and Development Center. Wetzel, Radtke, Parchman & Seymour, (1996), examined 50 students who were instructed over five days on the repair of fiber optic cable with a structured format of lecture, computer-based training, demonstrations, laboratories, homework reviews, and question and answer periods. Students were approximately, but not randomly, divided among two DL groups (DL local and DL remote), and a comparison (non-DL) group. The DL technology was VTT compressed over telephone lines. The scores on the course final exam were slightly higher in the comparison group (86% correct) than in the DL local (85%) and DL remote (80%) groups, but this difference was not statistically significant. Although it took students longer at the remote site to complete their lab assignments, there were no significant differences between the groups in terms of procedural errors, observer ratings of safety, quality of work, or objective errors.

An interesting technique for measuring performance was applied here. Students were required to conduct a splicing exercise with the fiber optic cable. If successful, light would be emitted at the end point of the cable. During this hands-on performance measure, students were required to display the fiber end point to the instructor over the two-way video arrangement. A shining end point resulted in a passing grade.

Wetzel (1996) performed an evaluation of a refresher course in celestial navigation. Students (n=279) across two DL groups (remote and local) and a comparison group were compared on performance, reaction measures, and amount of interaction as determined by an observer. There were no significant differences among the DL groups on students' homework scores, but students in the remote group scored slightly, but significantly, lower on their final examinations than students in the local group. When inequities in seniority status were

controlled in this data, students in the remote condition still scored 4% lower than those in the local site.

Performance Measurement in Training

In the above examples, performance was measured primarily through written tests, although there were some examples of measuring hands-on performance. Two additional studies of DL in Army settings provide unique insight into the issue of DL and performance. One illustrates a pedagogical misjudgment about DL and performance in the training of complex perceptual and cognitive skills for air traffic controllers. The second addresses the use of behaviorally anchored rating scales to assess the long-term effectiveness of distance learning on task performance.

Application of Cognitive Skills. This research illustrates some limitations of VTT for training tasks that have time-sensitive response demands. Such tasks require that the student, based on a learned set of principles and rules, respond quickly and accurately to specific situations. The relevant point is that the acquisition of certain skills require conditions of learning that include individual training with frequent feedback and sufficient practice spaced over time. These learning conditions might not be present in all DL environments. The research was conducted during qualification training for MOS 93C, Air Traffic Control Operator (Wisher, Seidel, Priest, Knott & Curnow, 1997).

The traditional classroom training is an 11-week course at the Army Aviation Center and School, Fort Rucker. The training consists of both a knowledge component and a real-time performance component. The course has six phases, four (the knowledge component) having a written knowledge test and two (the performance component) having a hands-on performance test. Two of the knowledge phases, fundamental tower procedures and general topics, concerned learning declarative knowledge and facts. The other two knowledge phases, tower academic and radar academic, concentrated on learning principles and rule sets for later application in the performance component of the course.

A distance learning version of the MOS 93C course was prepared by the ARNG in coordination with the Army Aviation Center and School. Since time for training is more restricted for the ARNG, the DL course was extended to 11 months. The course was delivered through satellite-based VTT, with the instruction originating from Fort Rucker downlinked to eight remote sites. The DL training addressed only the four knowledge phases. Students who were successful with the knowledge component were then enrolled in the performance component during a special two-week program at Fort Rucker.

A total of 77 soldiers participated in the research, n=32 in the DL treatment group and n=45 serving in a classroom comparison group at Fort Rucker. The results demonstrated no significant differences between groups in the four phases of the knowledge component as assessed through a learning measure (Level 2 of the Kirkpatrick model). The average scores on an exam administered by the FAA after the first phase, for example, were 88% for the classroom comparison group and 91% for the DL group (not significant). A knowledge retention test was

administered about 10 weeks after completion of the fundamental tower phase. The results indicated that the comparison group had a knowledge loss of 15% and the DL group a loss of 14%, right in line with expectations (Wisher, Sabol & Ellis, 1998). By these measures, both groups were equally successful in the acquisition of knowledge. The results for the hands-on performance phases, however, were markedly different.

The two hands-on laboratories, Tower Laboratory and Radar Laboratory, were conducted at Fort Rucker. Success on these phases required application of principles and rules acquired during the knowledge component, but now under time sensitive conditions. For example, in the Tower Laboratory, students were required to issue a radio call to an incoming flight. This depended on calling out appropriate phraseology for issuing advisories and control instructions, air traffic clearances, and taxi instructions, all learned during the Tower Academic knowledge phase. The completion rates for the classroom comparison and DL groups for both laboratories are presented in Table 2. (Note: The attrition that occurred throughout the 11 month DL course resulted in a sample too small for meaningful testing.)

Table 2. Completion Rates for Skill-based Performance

	Tower Laboratory	Radar Laboratory
Comparison Group	90%	85%
Distance Learning Group	58%	14%

It is apparent that the DL group had problems in applying the procedures and rules learned through the VTT program. Among the possible explanations for the poor performance is the effect of rearranging the training schedules to accommodate the ARNG's scheduling constraints. In contrast to the DL group, the classroom comparison group received the Tower Academic phase followed immediately by the hands-on Tower Laboratory phase. The Radar training had a similar ordering. In the DL version, students were trained on both academic (knowledge) phases before beginning the hands-on laboratory (performance) phase several weeks later. A retroactive interference effect could have occurred. Here, the memory consolidation of the rules and principles for one phase could have interfered with the consolidation of the other phase prior to their application. A second factor is the time delay (and knowledge decay) from initial learning to application, although review sessions were available. A third factor concerns the use of multiple-choice recognition tests to measure original learning, which is not compatible with the task requirement of rapid recall and action. The lesson learned is that the conditions of learning must be recognized and taken into account when converting a classroom course to a distance learning format. In this case, a computer-based program that simulated the Tower and Radar laboratories could have established a means to transfer the knowledge to the hands-on tasks without delay and without a potential retroactive interference effect. The instructional method embedded in a DL medium should be the determining factor in determining the appropriate delivery medium (Clark, 1994).

Distance Learning and Job Performance. **The U.S. Army Sergeants Major Academy (USASMA) is responsible for preparing noncommissioned officers for assignments as battalion and brigade staff NCOs. A four week course, taught either in residence at**

USASMA or through VTT, is required to obtain the additional skill identifier as a qualified Battle Staff NCO. ARI is engaged in a study (scheduled for completion in October 2000) that is examining the relative effects of the traditional classroom or VTT distance learning versions of the course on job performance. This is an example of the behavioral measure (Level III of the Kirkpatrick model).

Working with subject matter experts, behaviorally anchored rating scales were developed for each of eight performance dimensions. The performance dimensions were:

1. Assists in the military decision making process
2. Prepares combat orders or annexes
3. Prepares or constructs graphics or overlays
4. Understands the intelligent preparation of the battlefield
5. Assists in planning of Army operations
6. Assists in the planning and execution of CS and CSS
7. Manages record keeping
8. Prepares and conducts military briefings

A sample of approximately 400 soldiers who completed the BSNCO course, either the classroom or VTT versions, between February 1999 and February 2000 were selected for the study. Their immediate supervisors were identified, and the rating scales were distributed to them, along with a videotape of the Sergeant Major of the Army urging them to complete the rating form. The rating forms were to be completed by the supervisor after the selected NCO had between six and nine months of battle staff experience. In addition to this behavioral measure, school records yielded scores on four written tests during the course as a learning measure (Level II of the Kirkpatrick model). Finally, a course satisfaction survey was administered to students at the completion of either version of the course (Level 1 of the Kirkpatrick model).

Although the results will not be known until October, this study represents an important methodological practice of assessing performance on the job for eight separate dimensions. The method to create the multi-dimension performance measures was modeled after that developed for Project A, which over a decade earlier was employed in the revalidation of the Armed Services Vocational Aptitude Battery (Campbell et al., 1985). Such a methodology allows the relative effects of distance learning can to be compared, dimension for dimension, with a classroom comparison group. Certain tasks might lead to improvements in performance for one instructional method but not for others. The results of this study will be of particular importance to understanding the long-term effects of distance learning on performance.

In summary, military training taught through various distance learning delivery media, video, audio, computer mediated conferencing (CMC) and audiographics, have exhibited little improvement over conventional classroom instruction. The studies with VTT have been consistent across the services, but the audio and CMC technologies have had few applications reported. In view of the widespread availability of the Internet or military intranets to foster the Army distance learning vision of “anytime, anyplace” learning, the new e-learning tools

emanating through the World Wide Web use are discussed later in the section on future considerations.

Measuring Performance Using Technologies

Besides delivering instruction, DL technologies can be employed to measure performance. An example mentioned earlier was the use of the fax to transmit work samples (clerical forms) to a central performance assessment facility. Another was the use of video to transmit images of the outcome of performing a task (a shining fiber optic cable) to the instructor at the origination site. If used properly, a one-way video, two-way audio could be used to assess performance on observable tasks, such as most of the common soldiering tasks. The proper equipment for testing must be available at the remote sites and the video camera must be capable to track soldier movements. The feasibility of this method was demonstrated by the Pennsylvania Army National Guard during the evaluation of medical tasks during a DL pilot test of training combat lifesaver skills.

Another window to observe performance in DL is the use of audiographics to monitor performance while students are learning and practicing a digital task. A study was conducted on the production and delivery of a valid USMTF (United States Message Text Format) message (Freeman, Wisher, Curnow & Morris, 2000). A key enabling objective was to understand the composition of a message. This required the ability to identify the structural components of a message and to become familiar with the rules for structuring these components. Also required were an understanding of different message formats, occurrence categories, special use characters, and how to correct message errors. Since there are hundreds of message types, the hands-on portion of the training required that only a representative sample be executed during the training period. The hands-on portion of training was conducted individually on a personal computer linked to the Internet. Remote sites at Fort Hood, Fort Leavenworth, and a reserve center in Milwaukee participated in the one-day course.

Audiographics technology was used to enable a two-way, interactive replication of each learner's screen to a separate monitor at the instructor site. This was accomplished through the white boarding and collaboration sharing functions described in the T.120 standard. The instructors viewed a cluster of six monitors to independently view each learner's performance. Each cluster representing the students at a remote site. The monitors reflected student actions while attempting the digital skill (message composition) during a hands-on laboratory exercise. The instructors could assume control of each learner's application independently for demonstrating correct procedures. The instructors reported this capability as more effective than the practice in the conventional computer classroom: roaming about the room and peering at an individual's progress while giving verbal feedback when needed.

Instant Messaging. Another innovative feature of this study was the use of the instant messaging function of the T.120 standard. It was provided as a means for students to ask questions and for the instructor to privately coach /assist each student. Instant messaging (IM) is a relatively new Internet application that enables users to create their own private chat room and is now the preferred medium of immediate communication between users. In the Freeman et al.

(2000) study, IM was enabled between instructors and students such that the instructor assigned to observe performance remotely was able to "converse" independently with any student through a textbox. When students were having problems with the task, an IM was issued to the instructor. Also, when the student appeared to be stuck on a task, the instructor was able to issue an IM to that student as a means of timely performance feedback. A coaching dialogue ensued and the problem would quickly be rectified. Students (n=38) completing the three hour hands-on performance segment engaged 651 IM transactions. This rate of questioning and individual feedback is over 5 times the documented rate of questions in conventional classrooms (Graesser & Person, 1994). The Freeman at al. (2000) study was a pathbreaking application of the T.120 telecommunications standard for assessing performance in online training environments.

As described below, future directions in DL point to Web-based environments with greater emphasis on interactions between students. As Fetterman (1998) accurately points out, technology tools are playing an increasing role in e-learning research. There are now Web tools for data collection, analysis, and reporting. Technologies exist for recording online interviews, sharing data and resources, organizing field notes, searching database engines, locating needed resources, and analyzing discourse. Indeed, one might describe this as a revolution in assessment tools. Electronic surveys are also growing in popularity and usage (Champagne, 1998).

Future considerations

Training in the future is destined to be more soldier-centric, with the individual soldier assuming more responsibility for his or her learning. Soldiers will have more control of their learning along with more responsibility. On a broader level, there is an interest in developing multi-skilled soldiers for the Army Development System XXI Task Force, which calls for "adaptable" soldiers. This may require soldiers skilled in what are now considered separate battlefield functions or systems to possess also the capacity to adapt rapidly to changing situations, scenarios, missions, etc. Metacognitive abilities, peer mentoring, and collaborative learning will be relevant factors in preparing soldiers to be adaptable.

In the educational research and Web-based instructional marketplace, trends in pedagogy are converging with the emergence of e-learning technologies that allow for greater learner control, personal responsibility, and collaboration. These are in line with the Army goals towards a learner-centric model. The Army has initiated a Science and Technology Objective in FY 2001 on "Training Tools for Web-Based Collaborative Environments" which will seek effective ways to train and measure performance using Web-based environments. The prospects for adapting the new genre of Web-based tools from educational to training applications has been reviewed in detail by Bonk and Wisher (2000). Some of the considerations discussed in that report are summarized here.

E-learning is a unique context wherein learner-centered principles are particularly relevant as students become the center of the learning environment. In fact, in successful online courses, students might assume significant instructional roles such as offering instructional tips and constructing new knowledge that were once the domain of the instructor (Harasim, 1993). Along these same lines, Levin and Ben-Jacob (1998) predict that a key future component of

learning in higher education will be collaborative learning. Such student-centered learning environments will undoubtedly include team learning opportunities.

If the Army is to gain the full benefits of online instruction, a significant change in the preparation of instructors will be required. TRADOC plans call for instructors or mentors to be assigned to each learner in a DL course, including online courses. The lessons from education are that online learning is an entirely new type of pedagogical experience requiring a redesign of instructor roles, responsibilities, and commitments as well as support and training for those teaching online (Besser & Bonn, 1997; Doherty, 1998). The potential modifications in instructional roles might seem overwhelming. A summary of these are that the instructor will move:

- From information provider to facilitator guiding learning.
- From group instructor to one-on-one leadership role.
- From lecturer to co-learner participating in online activities.
- From platform pedagogue to online host, connecting learners for discussions and debate.

A year-long faculty seminar on online learning at the University of Illinois recommended that online instructors limit lecturing while monitoring and prompting student participation, organizing student interactions, and writing integrative and weaving comments on occasion. Until instructors are prepared and feel comfortable in these new roles, online courses may experience higher than expected attrition rates.

Online Learning Issues

The lessons being learned in the development and evaluation of online learning programs are emerging from higher education. Over 54,000 courses are now on line, and the professorial ranks are divided on the merits and threats of online learning. One byproduct of online learning in the military will include written products – plans, orders, recommendations, and decisions regarding operations on the digital battlefield -- developed by students. In the future, these discourse forms will serve as measures of learning and gauges of performance. They are the products of critical thinking and group problem solving, cast in the form of essays, emails, chats, and threaded discussions during individual and collaborative learning exercises. Clearly, qualitative and quantitative tools and techniques for measuring such discourse will be fundamental to the measurement of performance in these learning environments. Described below are some lessons learned, pedagogical practices, and evaluation methodologies for measuring performance in online courses. These should be considered carefully for future use in the Army.

Measurement Instruments. The literature on online learning details both quantitative and qualitative research instruments (Riel & Harasim, 1994). On the quantitative side, researchers often discuss usage patterns, computer log data, data mining, video screen grabs, participation rates, student and instructor attitudes, writing skill improvement, peer responsiveness, and various data mining methods. Data mining tools now enable researchers to quickly obtain basic or summary usage statistics, classification and association analyses, time-series analyses, and

data visualization depictions (Harasim, 1999). Such tools can elucidate the timing and quantity of student online work as an independent variable predicting a learning outcome.

Quantitative measures can also assess student skills or traits. For instance, with the heavy emphasis on writing and communicating in most online learning environments, it is not surprising that there is interest in writing skill development (Bonk & Sugar, 1998). Lexical semantic analysis, the development of which the Army is co-funding, is an advancement in the quantitative measurement of certain qualities of written discourse. It may be of service in assessments of written products devised by online learners, but it remains under investigation. Other automated measurement tools, such as the Project Essay Grade led by Page, reviewed by Hiller (1998), can supplement instructor feedback on grading student writing in DL assignments by providing summary statistical data on writing features, such as average word, paragraph, sentence and composition length. Hiller's work, initiated under PEG, departed by successfully employing content analytic techniques suitable for providing writers with feedback tied directly to their words and phrases. This methodology is suitable for: a) grading quality of compositions, b) scoring for content knowledge in short essay tests, and, c) of greatest importance for writing instruction, presenting feedback to writers on effective and ineffective word usage (e.g., spotting and encouraging the use of examples and illustrations, as cued for the computer by use of "for example," "to illustrate," "such as," "e.g.," and discouraging use of features such as double negatives, passive sentence constructions, etc.).

***Content Analysis*. The tools for assessment on the qualitative side are also rich and varied. Here, researchers often point to interaction and content analyses, discourse quality, verbal protocols, message flow analysis, message thread analysis, semantic trace analysis, forms of feedback, observation logs, retrospective analyses, and user think alouds. In fact, so many methods are mentioned in the literature, it is difficult to know when and where to use them. Message thread analysis entails grouping messages related to one another into common message threads for analysis (Riel & Harasim, 1994). Another qualitative technique, semantic trace analysis, is designed to map out the development of a single idea or set of ideas over time. Using this latter method, one might discover the source of pivotal student contributions (Riel & Harasim, 1994).**

Messages. Researchers point out that how often a message is referenced by other messages is an indicator of the importance of certain network participants and the direction of the online conversation. Graphic displays of message interaction might signify not only what topics were popular but also member status and dominance. Messages within a discussion thread might be classified according to whether it is in initiation of a discussion, a reply, or an evaluation. Noting who is performing such acts—instructor or student—is useful in determining whether the online discussion is following traditional instructor domination patterns or allowing for more student-centered learning.

As the e-learning assessment tools evolve, researchers might look at both quantitative and qualitative data with student questionnaires and related evaluations, performance measures, observations of interaction patterns, technology evaluations, completion and attrition rates, and cost-benefit analyses (Owston, 1999; Phelps et al., 1991).

Online discussion analysis. Curtis and Lawson (1999) designed a scheme for analyzing online discourse. They proposed greater understanding of the types of behaviors typically found in collaborative learning situations. Their coding scheme categorizes such high level behaviors as planning, contributing, seeking input, reflection and monitoring, and social interaction. As in other studies, few students challenged others or attempted to explain or elaborate on their particular positions.

Taking a more mathematical approach, Hara (2000) recommends Formal Concept Analysis (FCA) for understanding conceptual hierarchies in e-learning. FCA is based on a mathematical lattice theory that analyzes quantitative data visually. According to Hara (2000), it can be used to describe social relationships. For instance, she used it to reveal complex relationships among categories of coded data in online environments, thereby providing insights into online interactions. A simpler scheme was used by Hoffman and Elliot (1998) who coded Web dialogue according to the six levels of Bloom's taxonomy. They found that student Web electronic dialogue occurred at a deeper level than their more superficial written journals. These researchers concluded that case-based discussions on the Web could foster student problem solving, interaction, and the creation of a network of peers with whom to communicate. Such techniques may play a role in assessing performance in preparing multi-skilled, adaptable soldiers.

Shareable Courseware Object Reference Model (SCORM)

The SCORM is an evolving specification by industry and DoD to develop a standard for tagging learning content in Web-based environments. The tags are formatted in an Extensible Markup Language (XML)-based representation of course structure. They can be used to define all course elements, structure, and external references so that courses, or any of their elements, can be interchanged and moved from one learning management system to another. This would enable Army schools to freely exchange content from one training context to the other amongst other services and content providers. The savings in course development can be substantial, and the future prospect to create specialized "learning objects" on the fly to address specific training deficiencies of individual students is far reaching. Version 1.0 of the SCORM was issued in January 2000.

It is not clear whether performance specifications will be included in future versions of the SCORM. The addition of standard meta-data markings to indicate performance metrics would be useful, such as the Army's view of learning objectives as clearly and concisely describing student performance required to demonstrate competency in the material being taught. The extent to which this will be included in later versions is to be determined. It is an opportunity that must be explored.

Summary

The published literature in educational settings on the effectiveness of DL is overwhelmingly anecdotal. Evaluations are usually informal and conducted by users rather than third-party independent sources. As a result, large-scale evaluations have tended to focus on issues such as usability, equipment quality, and learner preferences, rather than learning outcomes. Tests of academic knowledge, rather than performance-based measures, are by far the more common outcome measure reported. The majority of studies are not supported by an adequate experimental design and rarely offer objective measurement of performance variables (Wisher & Champagne, 2000). What conclusions about performance can be drawn from evaluations of DL that are often performed as autopsies--conducted when the program is completed to see what went wrong?

The most complete documentation for measuring performance from distance learning comes from the military training literature. The findings basically demonstrate that electronic replications of the classroom also replicate the learning outcomes of the classroom, leading to no performance advantage. Moving from the classroom replication to more robust pedagogical approaches and media combinations enabled through Web-based approaches may break through the no-significant-difference barrier.

The no-significant-difference finding has become a longstanding tradition in distance learning, dating to radio-based training in the 1940's. The lengthy list of articles, over 300 in a recent count (Russell, 1999), that paraphrase "there was no significant difference between the distance learning and classroom comparison groups" makes one wonder about the strides that have been made in applying other forms of instructional technology. For example, effect sizes of .4 to 1.05 are regularly reported in the meta-analyses of computer-based instruction and intelligent tutoring system. In these studies, an emphasis was placed on instructional design considerations, such as response cueing, adaptive instruction, informative feedback, and strong intrinsic or extrinsic motivation, Reigeluth (1999). These studies identify methods of instruction and situations appropriate for those methods. For studies that focus on the medium of delivery, Clark and Solomon (1986) conclude "past research on media has shown quite clearly that no medium enhances learning more than any other medium regardless of learning task, learner traits, symbolic elements, curriculum content, or setting."

Most implementations of DL are oriented to group instruction and appear to replicate the classroom environment. There appears to be greater concern for increasing bandwidth, ostensibly to improve the technical qualities of an instructor image, rather than improving the quality of learning outcomes and performance. Unlike computer-based forms of instruction, many DL applications lack the instructional advantage of individual feedback, adaptive cueing, and self-pacing.

Higher education and internal industry efforts are leading the way in pedagogical advances and clever use of Web-based environments for learning purposes. The impact of these innovative applications of media and instructional design for improving a soldier's capacity to perform military tasks, however, will depend on their adaptation to a military setting. Here, the benefits on performance can be measured in more realistic environments with clearly defined

standards and better-controlled conditions. Hopefully, the Army will be able to take advantage of these advances as it moves to the learner-centric model of anytime, anywhere training.

References

Alliger, G.M., Tannenbaum, S.I., Bennett, W., Traver, H., and Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*, 341-358.

Besser, H., & Bonn, M. (1997). Interactive distance-independent education. *Journal of Education for Library and Information Science*, *38*(1), 35-43.

Bialek, H., Zapf, D., & McGuire, W. (1977) Personnel turbulence and time utilization in an infantry division. (HumRRO FR-WD-CA 77-11) Alexandria, VA: Human Resources Research Organization.

Bond, C., & Pugh, J. (2000) Aviation skills training in the Army National Guard. *Proceeding of the 16th Annual Conference on Distance Teaching & Learning*, Madison, WI.

Bonk, C. J., & Sugar, W. A. (1998). Student role play in the World Forum: Analyses of an arctic learning apprenticeship. *Interactive Learning Environments*, *6*(2), 1-29

Bonk, C.J., & Wisher, R. A. (2000) *Applying collaborative and e-learning tools to military distance learning: A research framework*. Technical Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bramble, W. J., & Martin, B. L. (1995). The Florida teletraining project: Military training via two-way compressed video. *The American Journal of Distance Education*. *9(1)*, 6-26.

Campbell, C., Campbell, R., Rumsey, M. & Edwards, D. (1985). *Project A: Improving the selection, classification and utilization of Army enlisted personnel. Development and field test of task-based MOS-specific criterion measures.* (ARI Technical Report 717). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Champagne, M. V. (1998). Dynamic evaluation of distance education courses. *Proceedings of the 14th annual conference on Distance Teaching and Learning, 14*, 89-96.

Champagne, M. & Wisher, R. (2000) Design considerations for distance learning evaluations. In K. Mantyla (Ed.) *The 2000/2001 ASTD Distance Learning Yearbook*. New York: McGraw-Hill.

Clark, R.E. (1994) Media will never influence learning. *Educational Technology Research and Development*, 42, 2, pp. 21-29.

Clark, R. & Salomon, G. (1986). Media in teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

Curtis, D. D., & Lawson, M. J. (1999). *Collaborative online learning: An exploratory case study*. Presented at the International Conference of Merdsa, Melbourne.

Doherty, P. B. (1998). Learner control in asynchronous learning environments. *Asynchronous Learning Networks Magazine*, *2*(2), 1-11.

Druckman, D. & Bjork, R. (Eds.) (1994) *Learning, remembering, believing. Enhancing human performance*. Washington, DC: National Research Council.

DUSD (R), (1999). *Department of Defense Strategic Plan for Advanced Distributed Learning,* Undersecretary of Defense (Readiness), Washington, DC.

Fetterman, D. M. (1998). Webs of meaning: Computer and internet resources for educational research an instruction. *Educational researcher*, 22-30.

Fleishman, E. & Quaintance, M. (1984) *Taxonomies on human performance: The description of human tasks*. New York: Academic Press.

Freeman, M., Wisher, R., Curnow, C. & Morris, K. (2000) Distributed digital skills laboratory: A virtual coaching environment for information systems training. *Proceedings of the Interservice/Industry Training System and Education Conference, Orlando, FL, December, 2000*.

Gordon, E. (2000) *Skill wars*. Boston. Butterworth-Heinemann

Graesser, A. & Person, N. (1994) Question asking during tutoring. *American Educational Research Journal*, 31(1), 104-137.

Hara, N. (2000). *Analysis of computer-mediated communication using formal concept analysis as a visualizing methodology*. Paper to be presented at the American Educational Research Association, New Orleans, LA.

Harasim, L. M. (1993). Networlds: Networks as a social space. In L. M. Harasim (Ed.). (1993). *Global networks: Computers and international communication*. Cambridge, MA: MIT Press.

Harasim, L. (1999). A framework for online learning: The virtual-U. *Computer*, *32*(9), 44-49.

Harsha, B. (2000) Online training at Sprint. In K. Mantyla (Ed.) *The 2000/2001 ASTD Distance Learning Yearbook*. New York: McGraw-Hill.

Hiller, J. (1998) *Applying computerized text measurement strategies from Project Essay Grade (PEG) to military and civilian organizational needs*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hoffman, D. H, & Elliott, S. (1998). *Web conferencing case dialogues: A supplement to traditional journal writing assignments in educational psychology*. Paper presented at the Midwest Association of Teachers of Educational Psychology.

Joy, E. & Garcia, F. (2000) Measuring learning effectiveness: A new look at no-significant-difference findings. *Journal of Asynchronous Learning Networks*, 4(1), 33-39.

Kirkpatrick, D.L. (1984). *Evaluating training programs. The four levels.* San Francisco: Berrett-Koehler.

Kiser, K. (2000) E-learning takes off at United Airlines. In K. Mantyla (Ed.) *The 2000/2001 ASTD Distance Learning Yearbook.* New York: McGraw-Hill.

Koble, K. & Bunker, E. (1997). Trends in research and practice. An examination of the American Journal of Distance Education, 1987 to 1995. *The American Journal of Distance Education*, *11(2)*, 19-38.

Levin, D., & Ben-Jacob, M. G. (1998). *Using collaboration in support of distance learning* (Report No. IR019267). Orlando, FL: Presented at WebNet 98 World Conference of the WWW, Internet, and Intranet Proceedings. (ERIC Document Reproduction Service No. ED 427 716).

OTA (1989). "Linking for learning" A new course for education. U.S. Congress Office of Technology Assessment (OTA-SET-430). Washington, DC: U.S. Government Printing Office.

Owston, R. D. (1999). *Strategies for evaluation web-based learning*. Presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Payne, H. (1999). A review of the literature: Interactive video teletraining in distance learning courses. Second addition. Atlanta, GA: Spacenet, Inc. and the United States Distance Learning Association.

Phelps, R. H., Ashworth, R .L., & Hahn, H. A. (1991). Cost and effectiveness of home study using asynchronous computer conferencing for the reserve component. Research Report 1602, Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Phipps, R., & Merisotis, J. (1999). *What's the difference?: A review of contemporary research on the effectiveness of distance learning in higher education*. Washington, DC: THE INSTITUTE for Higher Education Policy.

Riel, M., & Harasim, L. (1994). Research perspectives on network learning. *Machine Mediated Learning*, *4*(2-3), 91-113.

Reigeluth, C. (1999) What is instructional design theory and how is it changing. In C. Reigeluth (Ed.) *Instructional-Design Theories and Model*s (Volume II). Mahwah, NJ: Lawrence Erlbaum.

Russell, T. L. (1999). *The "no significant difference phenomenon*." Chapel Hill, NC: Office of Instructional Telecommunications, North Carolina University. http://cuda.teleeducation.nb.ca/nosignificantdifference/

Sherry, L. (1996). Issues in distance learning. *International Journal of Distance Education, 1(4)*, 337-365.

Simpson, H., Wetzel, C. D., & Pugh, H. L. (1995). *Delivery of division officer Navy leadership training by videoteletraining: Initial concept test and evaluation* (NPRDC-TR-95-7). San Diego: Navy Personnel Research and Development Center.

Sims, P. (2000) Satellite training takes flight. In K. Mantyla (Ed.) *The 2000/2001 ASTD Distance Learning Yearbook*. New York: McGraw-Hill.

Smith, P.C. (1976) Behaviors, results, and organizational effectiveness. In M. Dunnette (Ed.) *Handbook of industrial and organizational pscyhology*. Chicago: Rand-McNally.

Tobias, S. (1994). Interest, prior knowledge and learning. *Review of Educational Research*, *64*(1), 37-54.

Walsh, W. J., Gibson, E. G., Miller, T. M., & Hsieh, P. Y. (1996). Characteristics of distance learning in academia, business, and government. AL/HR-TR-1996-0012, Brooks Air Force Base, TX: Human Resources Directorate.

Wetzel, C. D. (1996). *Distributed training technology project: Final project* (NPRDC-TR-96-7). San Diego: Navy Personnel Research and Development Center.

Wetzel, C. D., Radtke, P. H., Parchman, S. W., & Seymour, G. E. (1996). *Delivery of a fiber optic cable repair course by videoteletraining* (NPRDC-TR-96-4). San Diego: Navy Personnel Research and Development Center.

Wisher, R. and Champagne, M. (2000) Distance learning and training: an evaluation perspective. In S. Tobias & J. Fletcher (Eds.) *Training and retraining: A handbook for business, industry, government, and military*. New York: Macmillan.

Wisher, R. and Curnow, C. (1999) Perceptions and effects of image transmissions during Internet-based training. *The American Journal of Distance Education*, 13(3), 37-51.

Wisher, R. & Priest, A. (1998) Cost-effectiveness of audio teletraining for the U.S. Army National Guard. *The American Journal of Distance Education,* 12(1), 38-51.

Wisher, R., Sabol, M. & Ellis, J. (1999) *Staying sharp: The retention and reacquisition of military skills and knowledge*. (Special Report 39) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wisher, R., Seidel, R., Priest, A., Knott, B., & Curnow, C.. (1997). *Distance learning over extended periods: the effects of knowledge decay*. Paper presented at the annual conference of the American Educational Research Association, Chicago, IL.

Evaluating Large-Scale Training Simulations

Henry Simpson
Defense Manpower Data Center
DoD Center, Monterey Bay

ABSTRACT

Objectives of the manual are to (1) provide guidance to help analysts design meaningful training effectiveness evaluations, (2) describe procedures for alternative methods of conducting training effectiveness evaluations, and (3) provide examples of training effectiveness evaluations that may be used as models to emulate. Chapter 1 (Introduction) describes the problem and issues, objectives, and method. Chapter 2 (Building an Evaluation Framework) explains why people conduct evaluations. Chapter 3 (Evaluation Methods) describes evaluation methods and provides examples of their application. Chapter 4 (Case Studies) describes well-documented evaluations: SIMNET/CCTT (Simulation Networking/Close Combat Tactical Trainer) and MDT2 (Multi-service Distributed Training Testbed). Chapter 5 (Evaluation Problem Areas) contrasts laboratory and field evaluations, discusses lessons learned from past evaluations, and critiques field evaluation practice. Chapter 6 (Procedural Guidance) identifies and summarizes published evaluation guidance. Chapter 7 (Evaluation Criteria) discusses how evaluation criteria differ depending upon evaluation method, for small- and large-scale evaluations, and depending upon evaluation perspective (training versus system developer versus modeling and simulation). Chapter 8 (Evaluation Framework) presents the evaluation framework in terms of evaluation principles and a description of the timing of evaluation events, their purpose, and relevant dependent variables linked to relevant examples and procedural guidance.

(Note: The report is published in two volumes (Volume I: Reference Manual and Volume II: User's Manual). To obtain a copy of this report, refer to DMDC Technical Report 99-05.)

Making the Case for Training System (CCTT) Evaluation

Stephen L. Goldberg
U.S. Army Research Institute
Simulator Systems Research Unit
Orlando, Florida

In 1985, the U.S. Army Science Board conducted a summer study on training and training technology. A key recommendation of that panel was to endorse the need for complete evaluation of the results of training and the use of those evaluations for the improvement of training. The Science Board stated that effective and efficient training requires explicit quantitative measurement. Measurement was dubbed "The Missing Link." The Science Board's report provided three reasons for systematic measurement of training. It is necessary to provide feedback to trainers and training designers and to provide "Return on Investment" (ROI) information to senior managers to guide expenditure of Army training resources. These same reasons also apply to measuring the effectiveness of training devices, simulators or simulations (training systems). Decision makers are clearly interested in the training outcomes training systems produce relative to their cost, and trainers and training developers need effectiveness feedback in order to improve training strategies and their product's performance.

The 1985 Science Board report contained a number of other interesting recommendations for Army training. The report noted an imbalance in the Army training community's focus. Army Schools' spent most of their resources on individual training, providing Military Occupational Specialty (MOS) training, to soldiers and officers. Training research had also mainly addressed individual training issues. The Science Board recommended a change in emphasis from the individual to collective training, the training of the Army's crews, teams, and units. They called for development of integrated training programs for units and for help to be available for commanders to use them. The Science Board recognized a need to develop methods to quantify measurement of the effectiveness of unit training. They also recognized that training devices, simulators and simulations for either individual or unit training were being fielded without rigorous training effectiveness analysis.

Trends in Army Training

Since 1985, trends in Army training have been consistent with the Science Board's recommendations. More attention has been focused on unit/collective training. The Science Board's recommendation may have had something to do with this, but more likely, the development of the National Training Center, followed by the Battle Command Training Center, Joint Readiness Training Center and Combat Maneuver Training Center created a greater influence. A rotation at one of the Combat Training Centers (CTCs) has become the capstone event in a unit's training. Preparation begins six months prior to a rotation. During this period, emphasis is primarily, if not entirely, on unit training. With the exception of the Battle Command Training Center, the CTCs provide training exercises on instrumented ranges. The

CTCs incorporate technology that allows the conduct of realistic force on force and live fire exercises. During the force-on-force exercises each vehicle is instrumented for location and equipped with the MILES laser tactical engagement simulation systems. Units receive feedback on what their performance during After Action Reviews (AARs) that are partially based on the movement and firing event data captured by the instrumentation systems, and partially on Observer/Controller observations.

Field training is the traditional way that the Army has trained its units for combat. In recent years a number of factors have limited the Army's opportunities to conduct field training. The end of the Cold War has significantly reduced military budgets. At the same time political and environmental realities in the United States and Europe have reduced maneuver areas and precluded many of the field training practices which occurred during the Cold War era. German farmers no longer are willing to put up with maneuver damage and high noise levels. The increased speed, range and lethality of modern weapon systems have also outstripped the safety fans of most ranges.

Fortunately, at the same time as opportunities for field training have been limited, simulation technology has opened up new ways to train combined arms teams in realistic and challenging ways. The rapid evolution of simulation technology has allowed the Army to move toward more use of simulation-based training in preparation for the field-training culminating at CTC rotations.

Simulation Networking

Simulation-based fire and maneuver training had its beginnings in the DARPA Simulation Networking (SIMNET) program (Thorpe, 1987). It demonstrated the capability of man-in-the-loop simulators to create a virtual battlefield on which meaningful collective training could be accomplished (Alluisi, 1991). SIMNET was developed to demonstrate simulator-networking technology within a training system. DARPA eventually fielded 256 simulators at ten sites across the Army.

SIMNET was built to meet an 80% fidelity level. Simulators represented only key controls needed to accomplish fire and maneuver tasks. Other controls were represented two dimensionally on wallpaper. SIMNET was not designed with training features normally found in simulators. It did not have an operator station, nor an After Action Review data capture capability. The low cost and low fidelity design goals of SIMNET limited the range of tasks and battlefield conditions it could represent (Burnside, 1990). SIMNET, however, demonstrated to the satisfaction of the Army's leadership that distributed simulations could be a powerful training tool for Army units. The Close Combat Tactical Trainer (CCTT) program was launched in the early 1990's as a follow-on to SIMNET that would fix many of its short comings and expand its capabilities and task coverage.

Close Combat Tactical Trainer

The CCTT is a collective training system in which armor and mechanized infantry units man full-crew simulators representing M1 tanks and Bradley fighting vehicles. CCTT is designed to allow platoons and companies to conduct unit training in a combined arms environment (Johnson, Mastaglio, and Peterson, 1993). CCTT is an Army ACAT II procurement program. Unlike SIMNET, system design, reliability, documentation, logistic support and planned improvements followed standard practices for government procurements.

The fidelity of CCTT at both the system level and the level of the individual components is superior to SIMNET. The CCTT simulation environment includes many more of the elements found on the combined arms battlefield. CCTT supports changes in time of day. Time of day or night can be specified or conditions can be set to change as the exercise clock advances. CCTT night operations include the use of flares and night vision devices. Weather effects such as fog and rain can be simulated, allowing training to take place under conditions of limited visibility. CCTT's visual system portrays weapon systems effects and battlefield obscurants.

Training in CCTT allows for segmenting of the battlefield to train on specified aspects of a mission that need practice (Goldberg, Mastaglio and Johnson, 1995). Each exercise does not have to begin in an assembly area. An exercise can be started with the unit at Phase Line about to initiate its attack. If the attack did not go well, it could be repeated as many times as necessary and each repetition would be conducted under the same conditions as the first. CCTT allows for realistic play of Combat Support and Combat Service Support. Resupply and maintenance times are played accurately. Finally, CCTT includes an After Action Review system capable of capturing and replaying events on the virtual battlefield.

Field Training vs. Simulation Training

SIMNET's original training philosophy assumed that training in a virtual training system would be just like training in the field, Thorpe (1987). Experience with SIMNET and later with CCTT has shown that distributed simulations are not just like training in the field. Field training and virtual training do overlap in the tasks that can be effectively trained in each environment, but the overlap is far from complete. There are many tasks that can be trained better in a virtual environment than on a range or maneuver area and others where the opposite is the case. CCTT training does not allow soldiers to experience adverse conditions and real world weather and terrain. Field training provides realistic conditions for one time and place. However, safety factors, instrumentation limitations, and environmental considerations put limits on field training. The virtual environment also limits performance of many tasks. For example, defensive positions cannot be accurately represented because virtual terrain at this time is not dynamic. Bulldozers cannot dig a defensive position in CCTT's virtual terrain.

Field training and virtual (CCTT) training are not equivalent. A direct replacement of training time in the field with training in simulators is not possible since each environment provides capabilities and limitations not found in the other.

Training Effectiveness-Return on Investment

In a 1997 Department of Defense (DoD) Inspector General (IG) Audit Report an argument is made that the field exercise, as a training medium, is widely accepted by military leaders. The DoD IG concludes that the effectiveness of field training has been validated over an extended period of use. The counter point and main theme of the report is that the effectiveness of distributed simulation based unit training has not been proven and needs to be. According to the DoD, IG Report, DoD has invested $1.6 billion in large-scale networked simulation systems without evidence that they are effective. The report recommends that policy and procedures for evaluating the effectiveness and cost-effectiveness of large-scale training simulations be developed.

The DoD, IG Audit Report stated the need for training system evaluation in forceful terms. It is only one of many calls for more rigorous testing of the effectiveness of large-scale networked simulation systems. In 1989 the Army Science Board reported on a Close Combat (Heavy) training strategy for the 1990s. In that report, the ASB identified as the area of greatest need, the development of improved techniques to assess crew and small unit performance objectively and in near-real time. These assessments could be used for 1) instructional feedback and 2) testing and evaluating the contribution of training devices such as UCOFT, SIMNET, live fire and other approaches such as embedded training and instrumented ranges to Army training. The ASB recommended that the Chief of Staff of the Army establish a single point of responsibility for the management and validation of training devices.

The General Accounting Office (GAO) recommended in 1993 that the Secretary of the Army ensure that all testing, cost analysis and training effectiveness assessments be completed and fully considered before decisions are made about full-rate production of CCTT. The report also discussed the Army's plans to integrate CCTT with traditional field training. The question of the right mix of simulation and field exercises is not known due to a lack of data on system costs and quantitative assessments of how much each system contributes to overall training.

Within the Army, the Training General Officer Steering Committee identified the need for training effectiveness data for CCTT in 1997. Later the same requirement was identified by the Deputy Chief of Staff for Operations and the Commander of the U.S. Army Training and Doctrine Command (TRADOC) (Gelling, personal communication). The audit and Army Science Board reports, and taskings represent the Government's need for data regarding the "Return on Investment" from the CCTT and other training systems. This data is needed to make trade-off decisions and consider further investment in distributed simulation.

There have been numerous "ROI" questions asked with regard to CCTT. Many of these questions apply to all training systems Questions asked include:

What is the appropriate mix of live and simulation training?
Can simulation-based training (CCTT training) substitute for field training?
If so, what are the cost savings from less frequent field training?
Can skills be acquired and sustained in CCTT?
How much time should units spend in CCTT?

Does training in CCTT transfer to improved field training?
Does training in CCTT transfer to improved combat performance?

These are difficult questions that have serious resource implications for both funding of training systems as well as field training. Most were raised early in the CCTT acquisition process. The Initial Operational Test and Evaluation (IOT&E) for CCTT was to be the source of data to address many of them (GAO, 1993).

CCTT Initial Operational Test and Evaluation

The Army agency responsible for the CCTT IOT&E was the Operational Test and Evaluation Command (OPTEC). Operational tests are designed to evaluate the effectiveness of new equipment when it is manned by soldiers performing its intended functions in the field. Operational tests of weapon systems evaluate the effectiveness, reliability and maintainability of the system. While testing is conducted in a military unit context, the focus is on the weapon system. Questions addressed in weapon system operational tests include: is it hitting targets at acceptable rates and ranges; is the doctrine for the systems employment workable; how often is maintenance required; does it meet its reliability objectives. Operational testing of training systems present challenges to test designers since they do not fit the weapon system testing formula.

The plan for the CCTT IOT&E called for comparison of the performance of units who had trained with CCTT versus those that had not. The training effectiveness criterion measure stated that the CCTT trained units perform no worse than the units trained by traditional methods. The performance of units in the field is an indirect reflection on the training system's impact on unit performance. There are many factors that can influence unit performance in a field exercise, and the training strategy used to prepare for the exercise may or may not be the key contributor. The CCTT IOT&E plan was to evaluate the reliability and maintainability of the simulation system, but less emphasis was placed on evaluating soldier performance in the simulators. This is unlike the direct evaluations of weapon system performance discussed above.

In 1994 while the CCTT IOT&E test plan was being developed, John Boldovici and David Bessemer published an ARI Technical Report which discussed previous attempts at evaluating large-scale networked training systems (namely SIMNET). The report described the problems with these "one-shot" empirical evaluations that limited the inferences that could be generated from their results. Boldovici and Bessemer (1994) outlined a number of advantages and disadvantages of empirical training effectiveness evaluations. As a major advantage they note that results of empirical evaluations have been used as evidence to:

1) Support inferences about the effect of training systems on training outcomes;
2) Justify budgets;
3) Comply with acquisition regulations requiring test and evaluation; and
4) Recommend ways to increase simulator-training capabilities.

The major disadvantage according to Boldovici and Bessemer is that empirical evaluations are usually performed with limited resources, which causes compromises in research

designs and test execution. These compromises produce results that cannot support valid inferences with respect to transfer of training in networked training systems to performance in the field. The evaluation flaws noted were:

1) Insufficient statistical power to demonstrate transfer differences;
2) Inadequate sampling, resulting in confounding training treatment with pre-test proficiency;
3) Inappropriate statistical analyses;
4) Inadequate controls, which confound the effects of uncontrolled variables with the training treatments; and
5) Failure to collect data needed to properly interpret transfer results or needed to indicate ways to improve the simulation and its use for training.

Boldovici and Bessemer (1994) made a number of recommendations regarding how CCTT should be evaluated to overcome the problems found in earlier evaluations of SIMNET. They recommended that CCTT should be evaluated as a system, in relation to its role in Army training. The evaluation process should be continuous over a significant period of time. They assumed that there would be little known about how best to train with the first version of CCTT and that improvements in strategies and methods of use would occur as the Army gained experience with the system. Continuous feedback from evaluations would provide the information on which to make changes to the system and disseminate lessons learned.

Boldovici and Bessemer's 1994 report influenced OPTEC's test design but only to a limited extent. In the end, time and resource constraints drove the test plan to a design that was very similar to those used earlier with SIMNET. It addressed test issues in three areas, training effectiveness, reliability and maintainability, and the functioning of the mobile CCTTs.

The IOT&E (Operational Test and Evaluation Report on the CCTT, 1998) took place over a seven-month period from late 1997 till May 1998. It was conducted in three phases. The first phase took place in the fixed site CCTT facility at Ft. Hood, Texas. Phase II involved testing of baseline and a treatment battalion at the National Training Center (NTC), Ft. Irwin, California. The third phase evaluated the reliability of the CCTT mobile configuration. Cost constraints limited to one battalion the number of units that would receive training on CCTT followed by a rotation at the NTC. Data was collected from other units who participated in CCTT training or participated in an NTC rotation, but just one battalion did both. The treatment unit trained over a ten-day period in CCTT. The training consisted of 1 day of orientation and 2 days of exercises for each company, followed by 4 days of training as part of a battalion task force. No platoon exercises were conducted (Operational Test and Evaluation Report on the CCTT, 1998). The treatment unit developed their own training strategy; the testers did not control it. Unlike other home-station units that trained in CCTT, the treatment unit did not use the structured training scenarios that were available to them. Hiller's paper in this report explains how such results are ideosycratic and therefore not replicable or generalizable.

Results for the treatment unit showed relatively poor performance in CCTT compared to units using the structured training scenarios. However, the treatment battalion's companies performed better at the National Training Center than baseline units who had not trained in

CCTT. The treatment unit therefore met the criterion of performing at least as well as the units training without CCTT. OPTEC did recognize that sample size limited the adequacy of the test and their ability to fully evaluate training effectiveness. IOT&E results, briefed during CCTT's Milestone III ASARC, noted that OPTEC felt that continuous evaluation is required to fully address training transfer. The IOT&E report states that the following questions were not addressed by the test:

- The amount of training within CCTT that transfers to the field;
- The optimal strategy or mix of CCTT exercises within the current mix of live and simulator training;
- The optimal length of time a unit should train in CCTT; and
- The identity of which tasks are best trained in CCTT and which are best trained by some other method.

Given the questions not addressed by the IOT&E and the limited nature of the conclusions with regard to training effectiveness that were drawn from the test, one can conclude that answers to the "Return on Investment" questions discussed earlier still need to be found.

Training in CCTT

SIMNET's developers felt that training in SIMNET should be as close to training in the field as possible (Thorpe, 1987). This is why SIMNET did not include exercise control or After Action Review capabilities. Soldiers and leaders were supposed to experience the virtual battlefield in much the same way they would real battlefields. Applying the Thorpe approach, after the exercise, discussion of what happened would be limited to the perceptions of what each soldier came away with and would not benefit from comparison of soldiers' experiences to ground truth. The history of SIMNET utilization has shown that this approach was not effective. One of the first improvements to SIMNET was the addition of a replay device. In addition, utilization rates for SIMNET were low when units had to prepare their own scenarios, In addition, utilization rates for SIMNET were low when units had to prepare their own scenarios. Development of structured training scenarios, for National Guard units training in SIMNET under the Virtual Training Program (Hoffman et. al., 1995), were well received by their training audience and proved to be a model for future training support packages for large-scale networked simulators. An After Action Review System, and development of structured training support packages are available in CCTT as fixes for training limitations of SIMNET.

CCTT has been fielded at Ft. Hood, Ft. Knox and Ft. Benning. Construction is underway for CCTT sites at Grafenwoehr, Germany, Ft. Carson, Ft. Riley and Ft. Lewis.
Soldiers are training in CCTT everyday. Contractor Logistic Support provides operators for Semi-Automated Forces, After Action Review Stations and Exercise Control Stations. Forty training support packages have been developed and distributed to support platoon and company training in CCTT (Flynn et. al., 1998). The packages follow a structured training approach that provides units with everything needed for them to execute training scenarios developed to exercise specific collective tasks. Units training in CCTT have the option of using the training support packages or developing their own scenarios.

A training innovation recently developed for CCTT is the Commander's Integrated Training Tool (CITT). CITT is a web-based computer program that provides Commanders with the means to tailor existing Training Support Packages to meet their needs or develop new ones. CITT is currently being fielded to units with access to CCTT and its use will be trained as part of officer training courses at service schools. The use of CITT will increase the number of structured training scenarios available for CCTT training as existing packages are modified. CCTT users' training strategies and methods contribute along with the quality of training tools such as the After Action Review system and scenario generation tools (CITT) to training effectiveness.

The second important reason for doing training evaluation identified by the 1985 Army Science Board Summer Study was to provide Trainers and Training Developers with feedback on how training strategies and tools are working in order to make adjustments and product improvements. With regard to CCTT training questions like the following can only be answered by effectiveness data:

- What strategies are producing the best results in training performance?
- How should CCTT be used with an overall Combined Arms Training Strategy?
- What are trainers and soldiers' opinions of structured training packages?
- How is CITT being used? What do Commanders think of it?

Long-term Evaluation Planning

The CCTT IOT&E left important "Return on Investment" and training effectiveness questions yet to be answered. Based on the Training General Officer Steering Committee's stated need for training effectiveness data, the TRADOC System Manager for Combined Arms Tactical Trainer (TSM, CATT), the U.S. Army Research Institute (ARI), and the U.S. Army Operational Test and Evaluation Command (OPTEC) organized an Integrated Project Team (IPT) to develop plans for a Long-Term Evaluation (LTE) of CCTT to address these questions. A long-term approach was chosen, based on Boldovici and Bessemer's recommendations and to allow for collection of a large enough pool of data from across the Army to generate reliable results.

The IPT met for the first time at Ft. Hood, TX in March 1998. Attending the meeting were representatives from the three agencies mentioned above, Project Manager, Combined Arms Tactical Trainer (PM, CATT), the Armor School, the Infantry School, Seventh Army Training Center, Office of the Secretary of Defense Directorate of Test and Evaluation (DOTE), and the TRADOC Analysis Command (TRAC). The Aviation School sent representatives to later IPT meetings. Each of the agencies shared the same overall goal of wanting to learn more about how the Army uses simulation, but each brought its own perspectives and specific information needs regarding the Close Combat Tactical Trainer. Over an eighteen month period and a number of meetings, the IPT developed a set of objectives, agreed on Measures of Effectiveness (MOEs) and Measures of Performance (MOPs), and developed an overall

evaluation plan for a long-term evaluation of CCTT. The IPT also developed a briefing that summarized the plan that for presentation to the Army's senior leadership.

The plan included four objectives. The first was to demonstrate that training in CCTT improves task performance in the training system. To meet this objective task performance data would be collected during a series of scheduled Follow-on Operational Test and Evaluations (FOT&Es) scheduled for CCTT. Improvements in collective task performance would be tracked for each unit over the course of the training they received in CCTT.

The second objective was to identify the factors and conditions for effective training in CCTT. This objective was intended to identify those training practices employed in CCTT that produced superior training performance and those that resulted in poor performance. Factors to be tracked included training strategy, amount of training time allotted to platoon, company and battalion level training, use of structured training support packages, utilization of the Commander's Integrated Training Tool (CITT), effective troop leading procedures, and others. This objective was to identify training methods that were working or not working and to provide feedback on the effectiveness of CCTT training tools such as the After Action Review System and CITT.

The third objective was to demonstrate training transfer from the CCTT environment to a field-training environment. Bessemer (1990) had tracked performance during the field-training portion of the Armor Officer Basic Course prior to and following the introduction of SIMNET. Bessemer was able to show that, following introduction of SIMNET, the Armor lieutenant's performance in field training slowly improved. Gradually they were able to take on more advanced tasks than earlier students had been able to attempt. The third objective's intent was to replicate Bessemer (1990)'s approach at Infantry School following introduction of CCTT into their Basic Officer Course.

The fourth objective addresses the training mix question. It would identify strategies for incorporating CCTT training into a unit's training program that would result in effective overall field performance. Units' performance at externally evaluated exercises, such as CTC rotations, is a function of among other things, the training they receive. The fourth objective would require collection of data on units' incorporation of CCTT into their overall unit training strategy. In addition to tracking CCTT use, data would also have to be collected on how other training events such as situational training exercises, constructive simulations, gunnery training etc. were utilized. The training events, their frequency and quality would be related to performance by the unit at external evaluations, either at home station or Combat Training Centers. The analysis would be similar to that performed in the Army Research Institute's Determinants of National Training Center Performance Project (Holz, et. al., 1994).

Objectives two and four are the critical objectives in that they require data collection on a routine basis over a long period. Data collection locations would rotate to a different site each year during the five-year evaluation. The new location would have had its CCTT facility opened approximately six months prior to the start of data collection. Moving the data collection to different sites would ensure that results reflected practices across the Army and did not represent findings unique to one location.

Under leadership of the TSM, CATT the LTE plan was briefed to senior Army leaders including the Deputy Under Secretary for Operations Research, the TRADOC Deputy Chief of Staff for Training, and the Director of Army Training, Office of the Deputy Chief of Staff for Operations and Plans. In each case, the senior leader expressed agreement with the need for further evaluation of CCTT and with the plan's objectives. The cost of the plan and its length were questioned. In the end while the need for an LTE was recognized, no funds were forthcoming, nor was a potential source of funding identified.

The last LTE briefing took place in December 1999. Since then the subject has continued to be discussed, but no further actions have been taken. Whether any further training effectiveness analysis of CCTT occurs will be determined by how training system evaluation competes against other Army unfunded priorities. Given the recent history of training evaluation in the Army, the likelihood of success is questionable. Agencies that conduct tests and analyses have suffered severe cuts in resources. School-based test boards, such as the Armor-Engineer Board and the Infantry Board, have been eliminated. Directorates of Evaluation within the TRADOC schools were the first directorates to be reorganized out of existence when the Army downsized. The TRADOC Analysis Command (TRAC), the agency with the mission to conduct post-fielding training effectiveness analyses has taken severe cuts in personnel and funding. The Army Research Institute incurred serious cuts (on the order of 50%) in its staff and funding. Its interests are in the methodologies employed in performing training effectiveness analyses. The Army Test and Evaluation Command (ATEC) performed the IOT&E testing of CCTT, but its mission is limited to the formal role of test and evaluation in the hardware acquisition process.

In conclusion, the Congress, the GAO, DoD, the ASB, the Army's own leadership has asked that CCTT demonstrate its "Return on Investment". At the same time, trainers and training developers need feedback on the training products and approaches being used within CCTT. It is clear that the CCTT program would benefit from a training effectiveness evaluation. Whether one occurs is open to serious question.

References

Alluisi, E.A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors*, 33(3), 343-362.

Army Science Board (1985) *Summer Study on Training and Training Technology*, U.S. Army Science Board, Washington, D.C.

Army Science Board (1989) *A Close Combat (Heavy)Training Strategy for the 1990s*, U.S. Army Science Board, Washington, D.C.

Bessemer, D.W. (1991). *Transfer of SIMNET training in the Army Officer Basic Course* (ARI Technical Report 920), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Boldovici, J.A. & Bessemer, D.W. (1994) *Training research with Distributed Interactive Simulations: Lessons learned from Simulation Networking* (ARI Technical Report 1006) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Burnside, B.L. (1991). *Assessing the capabilities of training simulations: A method and Simulation Networking (SIMNET) application* (ARI Research Report 1565), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Department of Defense Office of the Inspector General (1997) *Requirements planning and impact on readiness of training simulators and devices* (Audit Report 97-138). Arlington, VA: author.

Drucker, E.H. & Campshure, D.A. (1990). *An analysis of tank platoon operations and their simulation on Simulation Networking (SIMNET)*(ARI Research Product 90-22), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Flynn, M.R., Campbell, C.H., Myers, W.E., & Burnside, B.L. (1998) *Structured training for units in the Close Combat Tactical Trainer: Design, development and lessons learned* (ARI Research Report 1727), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

General Accounting Office (1993). *Simulation training*: *Management framework improved, but challenges remain* (GAO/NSIAD 93-122), Washington, DC: author.

Goldberg, S.L., Mastaglio, T.W., & Johnson, W.R. (1995) Training in the Close Combat Tactical Trainer, in *Learning without boundaries: Technology to support distance/distributed learning,* Edited by Seidel, R.J & Chatelier, P.R., Plenum Press, New York.

Hoffman, G.R., Graves, C.R., Roger, M.E., Flynn, M.R., and Sever,R.S. (1995). *Developing the Reserve Component Virtual Training Program: History and lessons learned* (ARI Research Report 1675), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Holz, R.F., O'Mara, F., & Keesling, W. (1994) Determinants of Effective Unit Performance at the National Training Center: Project Overview. In *Determinants of Effective Unit Performance: Research on Measuring and Managing Unit Training Readiness*, Edited by Holz, R.F., Hiller, J.H., & McFann, H.H., U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Johnson, W.R., Mastaglio, T.W., & Peterson, P.D. (1993). The Close Combat Tactical Trainer Program. *Paper Presented at the Winter Simulation Conference 9*3, Los Angeles, CA.

Director Operational Test & Evaluation (1998) *Initial Operational Test & Evaluation of the Close Combat Tactical Trainer*, Washington, D.C.: author.

TEXCOM Combined Arms Test Center (1990). *Close Combat Tactical Trainer (CCTT) force development testing and experimentation* (TCATC Test Report No. FD 0200, RCS ATTE-3) Fort Hood, TX: Author.

Training General Officer Steering Committee (1997). Meeting minutes, Washington, D.C. (April, 1997).

Thorpe, J.A. (1987) *The new technology of large scale simulator networking: Implications for mastering the art of warfighting*. In the Proceedings of the 9th Interservice Industry Training System Conference (pp492-501), Washington, D.C., American Defense Preparedness Association, November, 1987.

Perspectives On Validity

Andrew M. Rose

The American Institutes for Research[9]

Introduction: Validity as a Process

Many people may think of validation as a property of an object: "This test has validity." "This training system is valid." Similarly, many people consider validity only in terms of an external, criterion-related outcome: "This test has a validity coefficient of X when compared to actual performance on the job." However, most training specialists and other professionals define validation much more broadly[10]. Decision-making regarding training effectiveness relies upon a set of inferences about operational performance demands, training system requirements, training content, the skills, knowledge, and abilities of individuals, theories of learning, transfer, and skill retention, and relationships among these factors. ***Validation is the process of testing the viability of those inferences.*** Current standards hold that a strong validation program is one that builds and weighs evidence about each of the inferences that lead to a final decision[11]. As pointed out by Dr. Hiller in his read-ahead paper for this conference, validating individual inferences within the decision system (a) increases the probability that the ultimate outcome of the decision-making system will be accurate, defensible, and explainable, (b) allows the isolation of different decision-making components for systematic troubleshooting of the entire system, and (c) builds on existing theory and empirically-based knowledge—enhancing our understanding of the viability of the decision system. Further, by viewing validation of training systems as a process of finding, building, and documenting relevant evidence prior to and over the lifetime of its use—rather than solely as the result of a single crucial study—a more realistic and practical approach to validation can be adopted.

To emphasize and extend this last point, we stress that criterion-related validation, or comparison against an external standard, is one of many sources of validity evidence. The effectiveness of a training system is not simply a function of one diagnostic test result. Many factors are weighed and combined. Erroneous decisions could result from problems with

---

[9] Copies of this paper may be requested by e-mail to Arose@AIR.org.

[2] See, for example, American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association; and Johnston, M. V., Keith, R. A., & Hinderer, S. R. (1992). Measurement standards for interdisciplinary medical rehabilitation. *Archives of Physical Medical Rehabilitation, 73*, S-1-S-22.

[11] Benson, 1998; Benson, J. (1998). Developing a strong program of construct validation: a test anxiety example. *Educational Measurement: Issues and Practice, 17*, 10-22; Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

instrumentation or problems with methods of combining information to make decisions. Specifying inferences and organizing research evidence relevant to those inferences illuminate the strengths and weaknesses in the decision system. Criterion-related validity studies are difficult, expensive, or impossible. We believe that it is essential to first assemble evidence that the new instrumentation and decision-making processes are sound before attempting to employ criterion-related validation using an external criterion.

You have all heard and understand the differences among the various forms of validity, such as face validity, content validity, construct validity, and criterion validity. Rather than bore you with discussions about these formal concepts, in this paper I will focus on what we all probably consider the key inference we are faced with when evaluating training systems: the notion of ***predictive*** validity. What information can or do we use and how can we make inferences about the future, especially in situations where no criterion performance measures are available (e.g., actual performance in combat)? The correctness and accuracy of predictions of future events cannot be judged *at the time of prediction*; by definition, predictive validity cannot be an empirical determination until the actual future is observed. I believe that by examining the more general issues of prediction, we will gain a better perspective on the concept of validity as the testing of inferences that draws on information from a wide variety of sources.

Methodological Perspectives on Prediction

In this section, we present some basic considerations regarding the concept of prediction. Our main purpose is to clarify the basic principles of prediction: the limits and potential scope of prediction, and the methodological requirements for predictive systems.

General principles of prediction

Although predictive systems vary over a wide range—from astrologers and other expert systems to computer-based neural networks—there are a few principles that hold for all methods:

- It is generally easier and safer (i.e., with less likelihood of being proved incorrect in the future) to predict general trends than specific developments. In weather forecasting, for example, "It is likely to warm up over the next few weeks" is a safer prediction than "The temperature will rise by 14 degrees next Tuesday."
- It is generally easier and safer to predict over the near future than over the long term; long-range forecasts are inherently more problematic.
- The fewer and cruder the parameters of a prediction, the more manageable and safer it generally becomes; it is by and large easier and safer to forecast aggregated phenomena than particular instances. For example, it is easier to predict unit performance than the performance of a specific individual.
- The more extensively a prediction is provided with a protective shield of qualifications and limitations, the safer in general is the prediction.

To expand a bit on the last point: the indication of possibilities and prospects is usually safer than that of real and concrete developments; it is generally easier and safer to deal in

possibilities and probabilities, to construct scenarios and answer questions about what *may* or *likely might* happen, rather than to make specific and committal forecasts about what *will* happen. This is not meant to imply that predictions should only be made under conditions of certainty; rationally warranted predictions can quite appropriately be based on generalized patterns that are not absolute and exceptionless but, rather, merely statistical.

Predictions can be either *categorical* or *conditional*. Categorical predictions have the form "X will happen," or "X will not happen," where X is some particular definite occurrence or outcome. Conditional predictions have the form, "X will happen if Y does." Also, questions can be of both forms: "When will X happen?" is categorical, and "What will happen if X does?" is conditional.

Conditional prediction comes in two types: specific or general with respect to time. "When you give me money, I will buy a boat" is an indefinite, but specific prediction. If this is stated as a "whenever" proposition—whenever condition C is realized, result R will ensue—it becomes a universalized conditional prediction: "Whenever people say hello to Andy, he will smile at them."

Requirements for prediction

All of the rational processes for generating predictions are, in the final analysis, based upon using historic and current information and projecting discovered patterns into the future. Any sort of rational prediction will accordingly require informative input material that indicates that three conditions are satisfied:

- *Data availability*: Relevant information about the past and present can be obtained in an adequately timely, accurate, and reliable way.
- *Pattern discernability*: The data exhibit discernible patterns.
- *Pattern stability*: The patterns so exhibited are stable, so that this structural feature manifests a consistency that also continues into the future.

These conditions are necessary, whether we are attempting to predict the date of the next lunar eclipse, or the outcome of the next Presidential election, or the combat readiness of units.

A Survey of Predictive Approaches

All true predictive methods involve examining existing data to seek out established temporal patterns and then projecting such patterns into the future. Prediction relies on the existence of a linkage between input data and predicted outcomes. This linkage can be based either on explicitly articulated principles or on personal judgments that exploit an expert's background knowledge about the matter. Predictions of the former sort are 'formal' and 'inferential' in proceeding via formally rule-specified modes of reasoning, while those of the latter sort are judgmental or intuitive in proceeding via the unformalized processes of reasoning in the personal estimation of individuals.

In Table 1 below, we list the major types of predictive systems. We then briefly discuss each type.[12]

Table 1: Predictive Systems

Predictive Approaches	Linking Mechanism	Methodology of Linkage
UNFORMALIZED/JUDGMENTAL		
Judgmental estimation	Experts	Informed judgment
FORMALIZED/INFERENTIAL		
Trend projection	Prevailing trends	Projection of prevailing trends
Statistical analyses	Data patterns	Curve fitting, regression, statistical modeling, pattern recognition
Cyclical analysis	System patterns	Pattern recognition and classification
Analogy	Comparable patterns	Classification and attribution
Indicator correlations and causal experiments	Causal correlations; statistical comparisons	Statistical inference
Law– or theory-based derivation	Accepted laws and theories (deterministic or statistical)	Inference from accepted laws and theories
Model-based derivation	Formal models (physical or mathematical)	Analogical inference and reasoning
Emulation and simulation	System-wide models	Pattern recognition, computer generation

Unformalized, Judgmental Predictions

Basically, expert judgment predictive systems rest on the assumption that since we (as mere analysts) do not sufficiently understand how existing data can be linked to outcomes, and therefore cannot provide any sort of cogent account for why the predicted result will indeed occur, we rely on an expert. In its simplest form, an expert judgment system involves presenting data to the expert—for example, the results of a training exercise—and awaiting a prediction (e.g., "This score is high enough to convince me that the soldier will perform effectively in combat"). Expert opinions can be combined, either noninteractively (e.g., by averaging quantitative predictions) or through interactive consensus-formation processes embodied in collaborative groups or panels such as the Delphi method. Cooke (1991) has an excellent chapter on "Combining Expert Opinion" in his book, *Experts in Uncertainty: Opinion and Subjective Probability in Science.*[13]

12 This discussion draws heavily from Rescher (1998) *Predicting the future.* New York: State University of New York Press.

13 R.M. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science,* New York and Oxford: Oxford University Press, 1991, ch. 11.

An extension of this basic predictive system is to "extract" the relevant knowledge and inferential process employed by the expert so that we can model the "expert system"[14] in order to apply it to new situations. Prediction by means of expert systems uses experts as providers of information about reasoning and decision processes, rather than as mere data sources. The technique is to elicit and systematize the reasoning processes used by the experts; the idea is to elicit from the experts an inventory of the various factors they take into account and the rules they use to combine factors. Then, having used the experts as the source of methodological information, the process is systematized (usually via computer) and used to make prediction. Most prominently and successfully used in medical diagnosis,[15] expert judgment predictions have, in general, been good in engineering, fair in medicine, shaky in economics, and distinctly poor in sociopolitical affairs.

There is one fundamental flaw in this type of a predictive system: experts disagree. All too often we encounter situations that pit one judgment against another, and one "expert" against another. Only where the prospect of a direct reliance on scientific prediction is infeasible does the recourse solely to the predictive judgment of others make good sense.

Formal/Inferential Methods of Prediction

Trend projection. Basically, trend prediction consists of extrapolating future conditions from present ones. The simplest version of such a predictive system is simple linear extrapolation, based upon the notion that the future will continue the lines of past developments. Also in this simplest (and oversimplified) version, there is no attempt to understand the underlying causes of the trend.

The reason we have presented an oversimplified depiction of this method is that it represents a bridge between informal, judgmental systems and more formal systems. Most of our everyday predictions—what we expect to happen if we continue what we are presently doing—are essentially trend projections, although we perform no conscious 'linear extrapolations' when walking down stairs, when driving on a straight road, or when doing hundreds of other activities. However, linear trends seldom maintain themselves over longer periods of time; the road bends, we get a stone in our shoe. When these disturbances upset our predictions, or (to borrow a term from physics) we face nonlinearities in a dynamic system, we must resort to more complex predictive systems.

Statistical analysis. This class of predictive systems involves extrapolation of nonlinear—exponential, sinusoidal, S-shaped, etc.—patterns. Predictive analyses such as curve fitting can take a limitless variety of forms; there is an infinity of functions to choose from:

> Insofar as scientifically cogent forecasting involves the exploitation of patterns for predictive purposes, the tools of the statistician furnish a mainstay of predictive reasoning.…There are a great many powerful mathematical tools for fitting curves to temporally structured information: the data regarding the past can be projected by processes as varied as multiple regression analysis,

[14] An expert system should not be confused with an Artificial Intelligence (AI) system; in the former, we attempt to recreate the reasoning processes of a human expert, while in the latter we try to develop algorithms or heuristics, independent of how humans would approach the problem.
[15] R. Carande, "Expert Systems," Choice, 30 (1993): 1425-33.

time series analysis, envelope curve fitting, multimode factor analysis, correlational analysis, and various others. [16]

Since much of the existing statistical and methodological texts and references are devoted to this topic, and it is the area we are most familiar with, we need not discuss it further here.

Cyclical analysis. This type of predictive system is primarily used in economics and in some theories of acquisition of skill, where the underlying principle is as follows: Progress proceeds by way of successive spurts, with sequential stages of rapid development succeeded by longer exploitative phases, themselves leading to further periods of relative stagnation until the next triggering event. By determining where one is in the cycle, one can predict upcoming stages.

Notice that this approach is not quite the same as trend projection, statistical curve-fitting, or time series analysis, which depend upon extrapolation from current trends. Rather, cyclical analysis depends upon discerning historical patterns, isolating triggering events, and identifying indicators of individual stages. This method takes more of a holistic, theory-based conceptualization as the guiding predictive principle.

Prediction by analogy. Prediction often proceeds by drawing parallels between the pattern of events—either temporal or descriptive— in one domain to that of another, and then asserting that the pattern observed in the previous situation will apply to the current pattern. For example, the developmental course of one organism or nation or enterprise is often seen as a model available for use in guiding our expectations about others. The analogies usually are based upon descriptive similarities; they can also be based upon shared structures or common processes. One common method for prediction by analogies proceeds by placing a particular case into a statistical "reference class" of others with which it shares some salient features. For example, we predict longevity by placing an individual into a group identical (or highly similar) on age, medical history, weight, and other presumably important dimensions, and then making an actuarial prediction based upon the average age of the comparison group.

The strength of these types of predictions depends upon how closely the parallels can be drawn and the amount and quality of the information used to establish the parallels. The "tighter" or more well defined we can specify the reference group, the better the prediction. The main weakness of this predictive system is that analogies are never exact; furthermore, we cannot know *a priori* whether the parallel attributes we choose are the critical ones. For example, in the 1960s, forecasts of the development of modern power were made on the basis of growth curves observed for fossil fuel and hydroelectric power in the period from 1800 to 1960. These parallels and analogies missed several crucial dimensions and ultimately proved highly inaccurate.

Indicator correlations and causal experiments. A fairly common predictive method consists of generalizing from predictive indicators, based on empirical relationships (such as correlations) between observed factors or statistical association among variables in an experiment. Some examples include:

[16] Spirites et al. (eds.), Causation, Prediction, and Search. New York and Berlin: Springer Verlag, 1993.

- Training performance as a predictor of combat effectiveness
- Academic performance as a predictor of future earnings
- Medical symptoms as preindicators of the unfolding course of the disease
- Risk factors that predict a future disease
- The position and phase of the moon as a predictor of rising and falling tides

This procedure can take many forms. In general, such correlational predictions are based upon observed relationships, such as between smoking and lung cancer. Typically, a population is examined for potential correlations between two factors. A search is conducted, usually within a large statistical base, for interesting relationships. Inferences are made when X occurs with a different frequency among things having a particular property than among things not having that property.

While it is a basic truism that the presence of a correlation does not imply causality, we must be aware of some slightly more subtle problems:

Small sample size. It is always possible that an observed correlation is a coincidence; this is particularly problematic if we restrict our investigation to small sample populations. For example, suppose we find that a majority of subjects at a post who score higher than 80% on a test were from the southwestern states. Unfortunately, the sample size of this group is only 12. Chances are quite high I would not find an even distribution for all parts of the US; what we expect is some entirely expectable "clumping."

Incomplete data. It may be that the study and its results examined only part of the relevant data, or the study sample has inadvertently pruned away just enough data—say, by excluding certain subjects—to lend support to the idea of a correlation. Continuing the above example, I inadvertently failed to examine (or failed to report) the home states of other soldiers at the post. In fact, more soldiers there come from the southwest than any other area; while an interesting and suggestive phenomenon, this additional finding clearly weakens the causal hypothesis of a relationship between home state and success on the test. This example also illustrates the fact that correlations are frequently explained by some third factor that suggests a possible indirect link between the correlated factors.

A concomitant variation between A and B. Concomitant variation occurs when a variation in one factor, A, is accompanied by a variation in another factor, B. A likely explanation is that we have managed to pick two completely unrelated trends that happen to be going in the same direction at the same time.

Also falling in this category are causal experimental studies, where such indicator relationships are explored systematically. These experiments can take various forms, such as:

- *Randomized experiments*: A group of subjects are divided at random into experimental and control groups, and the suspected causal factor is administered to members of the experimental group only.

- *Prospective experiments*: Subjects are selected for the experimental group who have already been exposed to the suspected causal factor; control subjects are selected who have not been exposed to the suspected cause.
- *Retrospective experiments*: A group of subjects are selected, all of whom have the effect. These subjects are compared to another group, none of whom have the effect, in an attempt to discover possible causal factors.

When such causal experiments are conducted and inferences are drawn from observed relationships, we must be particularly careful in our interpretations. Certain experimental designs and statistical assumptions underlying the observed relationships limit the types of inferences that can be made.

Law- and theory-based prediction. A sophisticated predictive method is that of inference from formalized laws (generally in mathematical form) or well-established theories that govern the functioning of the system. For example, in astronomy we determine the requisite data regarding the present state of the system, then use physical laws to generate predictions about the system's future state. Much of physical science is based on derivations from quantitative "natural" laws. However, deriving predictions from qualitative relationships among variables is also quite common in other social science domains; for example, the rules of etiquette and rules of parliamentary procedure lead to reliable predictions of behavior.

Of more direct relevance, however, are the often-unstated presumptions of "laws" governing acquisition, retention, and transfer that presumably govern training. Some of these "laws" include:

- More training leads to better learning.
- Higher-fidelity training situations lead to better learning.
- Higher levels of initial acquisition lead to better retention.
- Higher levels of initial acquisition lead to better transfer to operational performance.
- Higher-fidelity training situations lead to better transfer.
- Higher levels of retention lead to better transfer.

While I believe that most training specialists believe that these statements are probably true in most cases, the points I want to stress are:

1. These statements are in fact derivations from *theories*, not from natural, universal laws. Each has been shown to be incorrect in certain circumstances.

2. The statements relating to operational performance are, at best, *predictions*; as such, they are subject to all of the vagaries of any prediction, such as the influence of unknown intervening events, unknown effects of time, additional learning that could affect previous learning and retention, and a whole host of other factors.

Model-based derivation. This involves using an artificially structured system to generate predictions. The system can be a physical model or a symbolic model (e.g., a system of differential equations). In this type of predictive system, we exploit the presumed structural

correspondence or "isomorphism" between the model and the system at issue to generate system predictions. Modeling by computer simulation is currently popular; it includes large-scale economic models of national economies, "virtual reality" simulations, and large-scale environmental simulations for weather forecasting. In their technical refinement, their precision, and their capacity to combine both scientific findings (natural laws) and the rules of thumb used in informed judgment ("expert systems"), computer models are the most flexible and powerful predictive tools we have. They have proven their worth in various areas, such as predicting population growth, stock market activity, vehicular traffic, atmospheric pollution, and national economic development.

The primary limitation of this predictive approach is that frequently there are inadequate data to support the operation of a workable model. Particularly for large-scale models, analytic complications arise because the real-world phenomena at issue are too complex. Similarly, the difficulty is that most large-scale forecasting models are based on a huge number of interrelated assumptions on which the predictive outcomes hinge. Variations in these assumptions can have a major impact on predictions. In the study of complex phenomena, it is difficult or impossible to establish the tenability of these assumptions. Nonetheless, these models are potentially powerful predictive tools.

Simulation and emulation. A simulation is a representation of a situation, person, group, environment, or any other system. The simulation includes what the researcher (constructor) considers to be the critical aspects of the system that are (or could be) causally related to the outcome of interest. For example, computer simulations of the environment used in weather prediction contain representations of ground, water, temperature, humidity, and numerous other variables or structures, including feedback loops. Simulations are basically descriptive models; they may attempt to represent the entire system or just some of the system components. My electronic keyboard simulates the acoustic properties of a piano: the simulation matches the pitch, loudness, and timbre of a piano. However, the simulation does not extend to the physical properties of a piano; as far as its appearance, the size and shape of the keys, the feel of the keys, key weight, damping, etc., the simulation is an inaccurate representation. Simulations can also be dynamic in the sense that the representations can include action. If I press the keys of my simulated piano, the keys will move. Weather simulations can be constantly changing as a result of initial conditions and feedback.

Emulations are predictions of what a simulation will do as a result of a specific input set. I can see what my simulation of the environment will do if the amount of carbon dioxide is increased; I can emulate system behavior for a variety of input conditions. Emulations are instantiations of (usually complex) predictions. We borrow the term from computer usage, where an emulator is a mechanism that allows a program written for one computer or platform to run on another system; the notion is that the system (or representation) will behave in a new environment or new set of circumstances not originally within its design.

A similar concept is projecting *scenarios*—pictures of alternative futures (possible and plausible). Developed out of wargaming techniques, this methodology has been applied extensively to issues in economics, politics, and international relations.

Perspectives on Evaluating Predictions

All of the rational processes for evaluating and validating predictions are in the final analysis based upon matching predicted behavior to observed events. Any sort of rational prediction requires input that meets three conditions:

- Data availability: Relevant information about the past and present can be obtained in an adequately timely, accurate, and reliable way.
- Pattern discernability: The input must data exhibit discernible patterns.
- Pattern stability: The patterns so exhibited must be stable, so that the observed structure demonstrates a consistency that also continues into the future.

While these may be necessary conditions, they do not guarantee a successful or useful prediction. Several additional dimensions should be considered in evaluating predictive systems, including:

- Importance of the prediction
- Detail or precision of the prediction
- Correctness and accuracy
- Credibility and evidentiation
- Robustness and convergent validity
- Reliability
- Generalizability

In this concluding section, we briefly describe each of these dimensions.

Importance of the Prediction: Is the prediction a useful piece of information? Predicting the course of the approaching hurricane is more valuable than predicting the course of barometric pressure readings. Predicting operational performance is more important than predicting training grades. Importance can be practical (e.g., important to understanding what event will occur or when it will occur) or theoretical (e.g., important to understanding why an event occurred).

Detail/Precision of the Prediction (specific vs. vague, particular vs. general, precise vs. imprecise): while one can "improve" the accuracy of a prediction by avoiding detail—thereby increasing its generality, vagueness, and imprecision—lack of detail decreases the value and utility of the prediction.

Correctness (true vs. false) and **Accuracy** (closeness to the truth): Predictions can be wrong in different ways. They can be outright wrong ("It will rain tomorrow"), they can lack completeness (e.g., not specifying all relevant contingencies), or they can be approximately correct ("It will rain approximately 2 inches tomorrow" when 1.9 inches fall). This criterion is tied to the above: how detailed or precise the prediction is. The more precise the prediction, the more likely it will be incorrect. Correctness and accuracy can be judged absolutely (How well does the system perform at correctly answering questions to which we know the answers?) or

comparatively (How well does this system perform in comparison to other systems in the same predictive situation?).

Credibility/Evidentiation: To repeat a statement made previously, the correctness and accuracy of predictions of future events cannot be judged at the time of prediction. While we can wait until some later time to judge accuracy, at the time of prediction the critical evaluative dimension is credibility: the justification for the prediction. When a prediction is made, we want to know why we should accept it. (Conversely, an accurate prediction without credibility is at best problematic.) Evidence supporting credibility can be direct—the correct prediction is made, along with an explanation—or indirect, by referring to the reliability and credentials of the predictive system. In the latter case, while we might not understand why a predictive process works, we must have reasonable assurance that it works.

Robustness/Convergent Validity via agreement with the indications of other predictive resources: A predictive system that generates forecasts that agree with other methods lends credibility to the system.

Reliability of the predictive system in terms of consistency of output across users, consistency of prediction given the same inputs on different occasions, and reliability of methods used to determine inputs (e.g., choice of predictor variables).

Generalizability or versatility as determined by the extent of the range of situations over which the system can function.

Conclusion

In sum, I hope that this brief introduction to issues surrounding prediction helps to broaden your perspective on some aspects of validity. I will have achieved my goal if you consider validation of a training system not as the result of a single experiment or field study, but as a gradual, and, if possible, systematic process of accruing evidence regarding the inferences you make from the data obtained.

Strengths and Weaknesses of Alternative Measures: Rating by Direct Observation, Objective Scoring of Results, Self Appraisal, Peer Appraisal, & SME Judgment

Larry L. Meliza
US Army Research Institute
Simulator Systems Research Unit

Timeliness of the Topic

The strengths and weaknesses of various means of measuring unit collective performance are timely topics as Army units experience force modernization. Units are being equipped with new and emerging: weapon systems; digital systems and; reconnaissance, surveillance, and target acquisition (RSTA) systems. Collectively, these systems are expected to improve battle outcomes through a variety of intermediate mechanisms. For example, digital and RSTA systems are expected to improve performance by increasing awareness of the tactical situation. The process of developing tactics, techniques, and procedures (TTPs) for the modernized force requires the ability to measure the presence and influence of these mechanisms (see Figure 1). For example, do the TTPs enhance situational awareness and the ability of units to make use of this increased awareness?

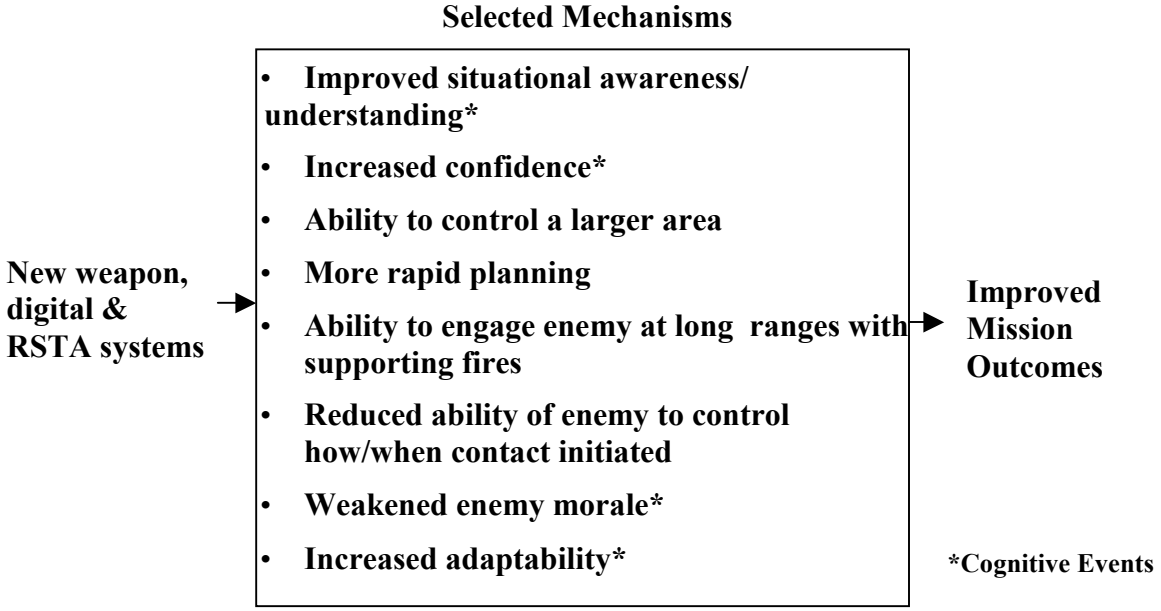


Figure 1. Force modernization measurement challenge.

The performance measures addressed by this paper include product-oriented and process-oriented cases. Process measures are viewed as being more diagnostic than product measures. A unit can do all the wrong things and yet succeed in terms of product measures, or the opposite case may occur. The current paper assumes that a degree of product measurement is required to support diagnosis. For example, have the processes employed by a unit resulted in the product of

increased situational awareness? Fortunately, as illustrated in Figure 2, many measures are both product and process oriented, depending upon one's perspective. For example, "whether an attacking unit engages the enemy at long ranges with supporting fires before the main body makes contact with the enemy" may be considered as a process-oriented measure that helps to explain the low number of casualties sustained relative to those inflicted (i.e., a product-oriented measure). On the other hand, one could also consider that this early use of supporting fires is a product gained, in part, from the process of "distributing fire support graphics in a timely manner."

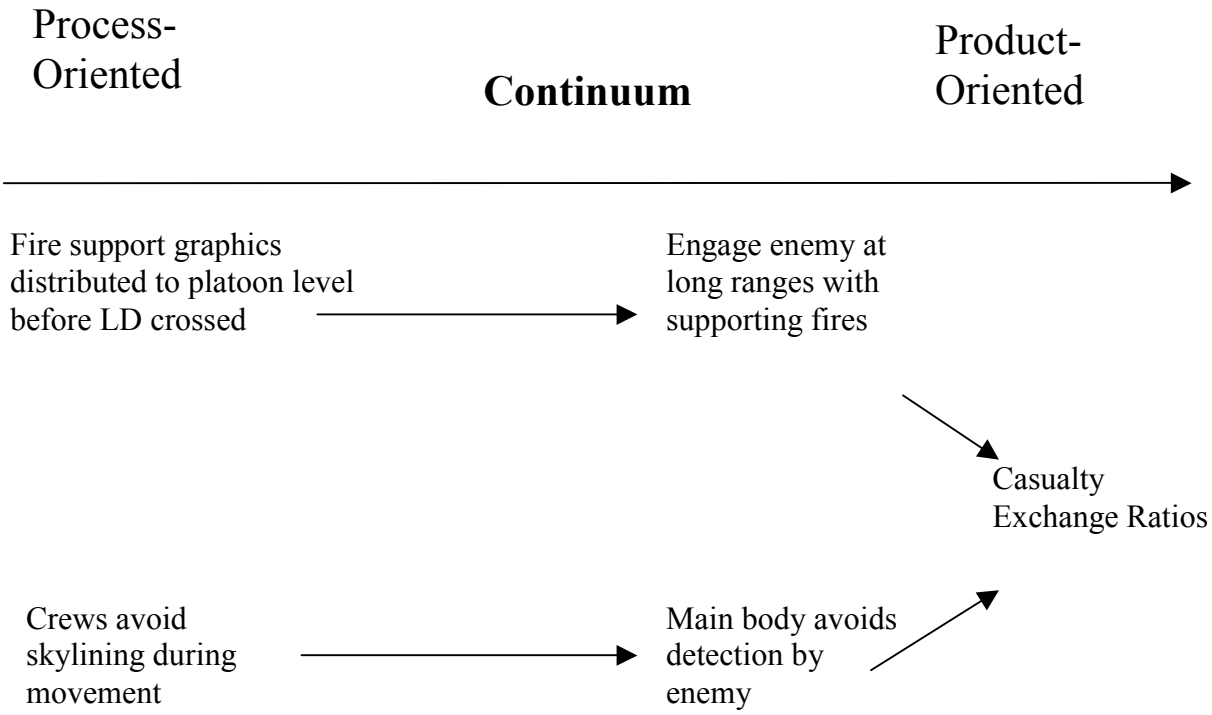


Figure 2. Many measures of performance can be employed as process or product measures.

This paper assumes that measures of performance vary along a continuum that ranges from highly process-oriented to highly product-oriented. Measures with a moderate degree of the product component are crucial in developing and refining TTPs because they measure the mechanisms by which force modernization is expected to influence unit performance. Advances in training technology have provided the Army with a rich electronic data stream describing exercise events on a moment by moment basis and tools to employ this data stream to illustrate key exercise events (Morrison and Meliza, 1999). These advances in training technology enable the Army to measure simulation events (e.g., are enemy forces engaged by supporting fires before the main body of ground forces is in contact with the enemy?), but the Army also needs the capability to measure cognitive events (e.g., is situational awareness improved? Are the confidence and morale of the enemy reduced?) It is crucial that we be able to "get inside the heads" of friendly and enemy soldiers and measure these cognitive "products."

Types of Measures

The five type of measures that can be applied in measuring the performance of units are described below.

Direct Observation

Direct observation of the behavior of leaders, soldiers, vehicles, and units is used to decide whether:

- the events promoted by tactical doctrine occur in practice (e.g., after issuing the OPORD, the leader asked questions to make sure everyone understood their roles and responsibilities)

- events that should not occur do/do not occur in practice (e.g., crews repeatedly engage enemy vehicles that have already been destroyed)

Observations may be guided by checklists during exercises, or they may employ displays prepared through the application of automation. Figure 3 illustrates how multiple measures of direct fire employment can be addressed by a single display. This display can be used to find out whether there are portions of the enemy force not being engaged by the friendly force, and it can be used to find out to what extent a unit is engaging enemy vehicles that have already been damaged/destroyed.

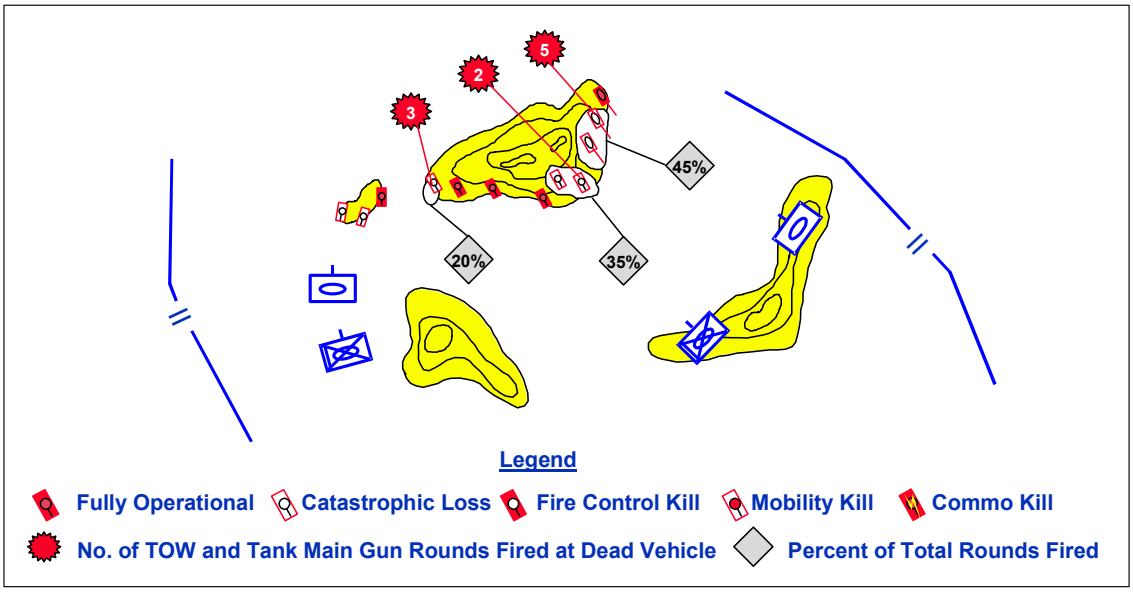


Figure 3. Display illustrating selected fire control elements.

Direct behavioral observations can also apply to mission planning and preparation activities at the command and staff level. Employment of the various Army Battle Command System (ABCS) components under the rubric of digitization enables automated information

displays that can be used to measure selected aspects of command and control activity (Gerlock and Meliza, 1999). For example, a display could be automated prepared showing when a battalion received its OPORD versus the time subordinate company teams receive OPORDS from battalion to find out whether the 1/3-2/3s rule is being applied.

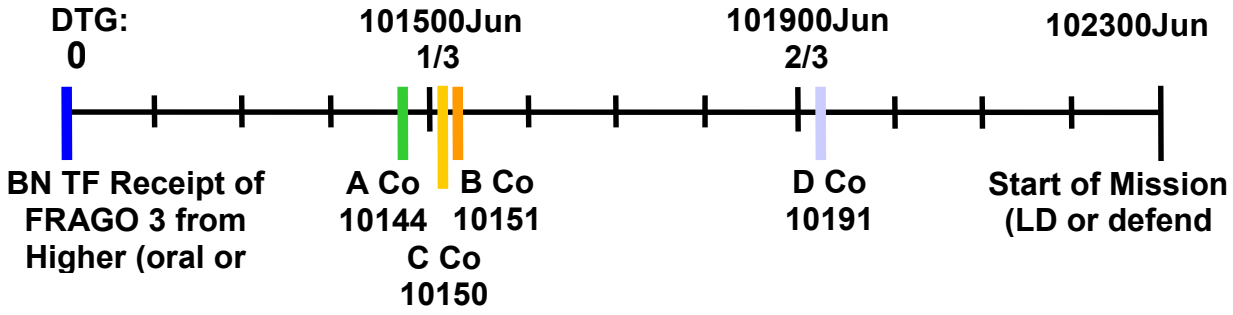


Figure 4. Time when each company team receives OPORD relative to time when battalion receives OPORD and time when unit should be prepared to perform mission.

Behavioral observations are intended to be very objective, but specific measures differ in terms of the extent to which they allow for subjective interpretation. For example, asking observers to decide whether all vehicle commanders are present when the platoon operations order is issued leaves little room for interpretation. On the other hand, asking observers to decide whether the platoon operations order addressed all mission, enemy, terrain , troop and time (METT-T) variables does leave room for interpretation. A knowledgeable observer may conclude, for example, that the platoon leader failed to mention an important interaction between the time and terrain situations.

Individuals responsible for observing behaviors should receive training on how to apply specific observable measures. Past research has shown that even subject matter experts (SMEs) need to be trained on their observation requirements to increase the reliability of the observations (Dwyer, Fowlkes, Oser, Salas and Lane, 1997). Unfortunately, the time and resources needed to conduct observer training are often inadequate.

Objective Scoring

Objective scoring involves an aggregation of behavioral observation data to provide a summary of some aspect of unit performance. For example, one might make a list of five events expected to occur when a leader issues an operations order (e.g., all subordinate leaders should be present) and a unit may receive a OPORD delivery score based upon the number of events observed. If the unit receives a low score, it indicates that practicing the collective task of OPORD delivery/receipt is warranted. Another, example of objective scoring would be to find that three of five indices of fire control problems were observed (e.g., less than half of the enemy vehicles in contact with the platoon were engaged, half of the platoon vehicles failed to engage the enemy, the platoon did not return fire within one minute of the initial enemy engagement) indicating a need for exercises in which fire control tasks are practiced. If scores indicate a unit is doing well in an area, even though it does not meet all of the criteria, practicing the entire task is probably not necessary.

Rarely will a single information display be able to provide all of the information needed to summarize a major aspect of unit performance. For example, there are important points of unit fire control not addressed by Figure 3 (e.g., are all of the crews and units involved in the engagement?) For this reason, an observer may need to examine multiple information displays to decide how many criteria were met regarding certain aspects of performance.

Peer Appraisals

Peer appraisals are judgments made by the opposition force or by higher, adjacent and supporting units regarding unit performance. Peer appraisals focus on how a unit's performance impacts performance of the peers. For example, the opposing force may conclude that a unit made poor use of cover and concealment, and it may or may not include data to justify the appraisal (we saw ten vehicles from the company team before it was in range of our weapons, giving us the opportunity to reposition our forces and maximize our firepower). Judgements made by adjacent, supporting, and higher units are likely to be concerned with describing problems encountered in working with a unit (e.g., Company A was not ready to make use of our engineering assets when we arrived and this resulted in a waste of our assets and time), and it may include data to justify the appraisal (we were at their position for over 30 minutes before we were given our first tasking).

If the data used to justify an appraisal are collected via a checklist, then this is simply another example of direct observation of behavior. If "peers" (enemy or friendly) provide the substantiating data on their own initiative, then it is part of a peer appraisal. The distinction is important to the individual responsible (trainer/tester/researcher) for implementing a data collection plan. In the first case, the data collection leader is responsible for distributing data collection forms and making sure peers understand how to apply the checklist items. In the second case, the data collection leader will probably need to interview peers to collect information on substantiating data.

Peer appraisals may be open-ended or closed. A data collector may ask the opposition force to rate or describe specific aspects of unit performance, or the data collector may simply ask the OPFOR to describe what they considered to be the most critical features of the performance of a unit. Data collectors may even ask the OPFOR how well they were able to estimate the courses of action taken by a unit.

An interesting example of the types of information that can be obtained from an OPFOR comes from the Force XXI Advanced Warfighting Experiment (AWE). In this case the OPFOR described how the mere presence of unmanned aerial vehicles tended to disrupt their mission preparation activities.

In some cases, peer appraisals (and self appraisals described below) may be the only source of information about a particular unit strength or weakness. One of the frequently observed, general types of problems in unit performance is a lack of compatibility among the standard operating procedures (SOPs)of subunits. An outside observer may not be aware of these problems, because the observer might not know about the SOPS, many of which are likely

to unwritten and informal. Another frequently observed problem is the lack of a common understanding of the task to be performed, especially when task performance requires different kinds of units to work together. Again, an observer may not observe the lack of common understanding, because it is a cognitive event.

Self Appraisals

Self appraisals may be provided by exercise participants when they estimate their own strengths and weaknesses (e.g., we took too long to select our firing positions). They may also provide data to substantiate the appraisals (e.g., by the time we selected our firing positions there was no time left to perform a reconnaissance of routes to alternate positions or finish preparing our positions). They can also provide information that helps identify causes of problems (e.g., we spent too much time planning before we started to prepare for the mission).

If an after action review (AAR) leader was aware of a unit's assessments of its strengths and weaknesses prior to the AAR, then this information could be used to reduce the AAR preparation workload and expedite the AAR process. For example, if the unit knew that its direct fires were largely uncontrolled during an engagement, there would be no need for the AAR leader to carefully guide the unit to this conclusion. Instead, the AAR could immediately focus on corrective actions.

Exercise participants can help explain what happened during an exercise by providing information about their thought processes. They can describe their situational awareness/understanding at various points during the exercise (e.g., we were able to verify enemy locations before we reached Phase Line Dog). Leaders may also provide information regarding what they did to monitor the performance of subordinates (e.g., I told my tank commanders to let me know when they had finished briefing their crews and checking their vehicles). Exercise participants also have information about unit SOPs and their impact on unit performance that is not readily available to an outside observer. As mentioned earlier in this paper, measuring these cognitive events is crucial in testing and refining TTPs for the modernized force.

A problem with using self appraisals is that they can be tainted by the unit's limited view of exercise events and by the desire to look good. Mirabella, Siebold, and Love (1998) have suggested two methods for improving the value of self appraisals. One approach is to use the Delphi method by having the various unit leaders rate their performance in an iterative fashion. The second approach is to compare self appraisals with the SME appraisals of unit performance.

SME Appraisals

An SME appraisal, like a peer or self appraisal, is a judgment regarding unit performance, but is it not limited to judging the impact of performance on the person making the appraisal. This particular class of appraisals also assumes that substantial expertise may be required on the part of the individual making the appraisal. For example, it is assumed that the individual evaluating a unit is capable of considering the impact of the specific METT-T

situation on performance. The resulting appraisal may summarize a key aspect of performance (unit fires were largely uncontrolled) and may or may not include a justification for the appraisal.

A recent ARI study considered the possibility of using SME appraisals to guide the AAR preparation process and other performance measurement activities (Brown, Nordyke, Gerlock, Begley II, and Meliza, 1998). Under this concept, a user can select from a menu of frequently encountered performance assessments. For example, the menu in Figure 5 shows unit maneuver performance problems from the Center for Army Lessons Learned (CALL) National Training Center Trends Analysis for the third and fourth quarters of FY 96. Each menu option can be used to call up a group of AAR aids that can be used to decide whether the appraisal fits a particular unit. For example, the aid shown previously in Figure 3 is one of those that can be used to decide if a unit "does not understand the fundamentals of direct fire planning." The great utility of implementing this concept is that it can be based upon the line of reasoning (here are the most commonly encountered performance appraisals and here are the information displays I would use to confirm or deny the correctness of each appraisal for a specific unit) used by the most experienced SMEs (e.g., observers/controllers with years of combat training center experience) and used by less experienced SMEs.

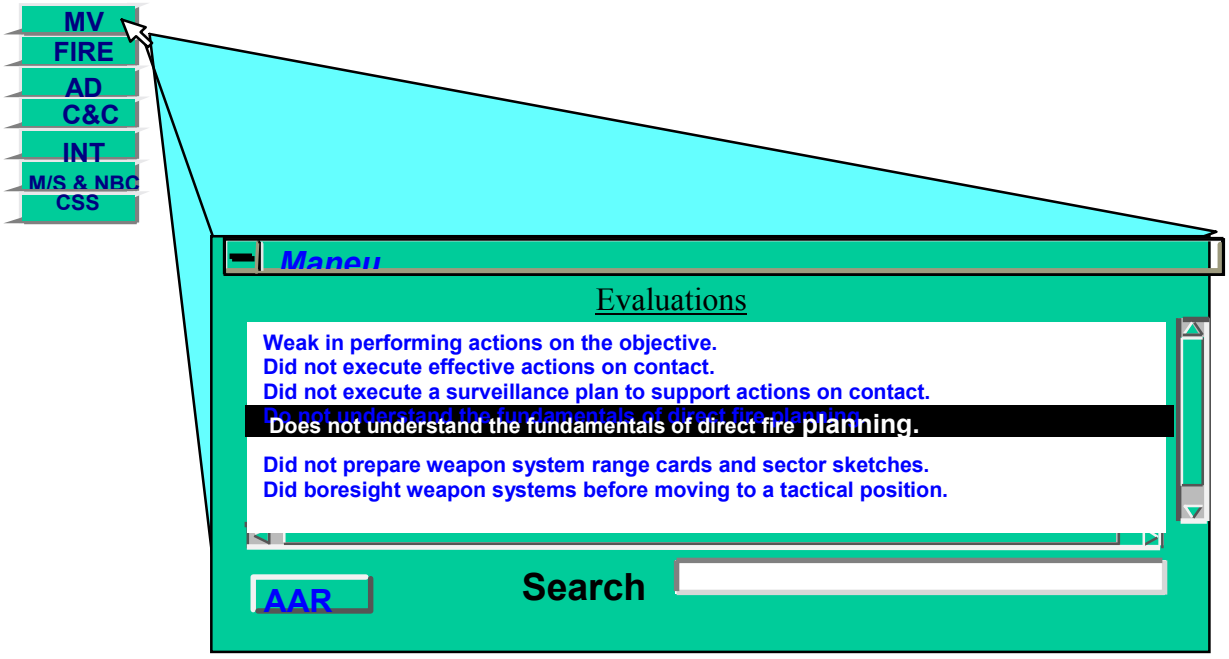


Figure 5. Use of SME judgments to trigger the production of candidate AAR aids.

Comparison of Measures

The six variables considered in comparing the utility of the five types of measures are listed and described below.

- Validity Does the measure do a good job of predicting future performance and does it measure a credible, meaningful capability?

- Reliability If the measure is applied by more than one individual will it tend to provide the same results for a specific instance?

- Workload Does application of the class of measures impose a heavy workload on trainers?

- Documentation Can it be used to illustrate/prove the existence of a performance problem?

- Corrective Action Does it point the way towards a specific corrective action?

- Sole Source Are their capabilities that can only be measured with a specific class of measures?

Reliability

Direct observation of behavior and objective scoring should be the most reliable measurement methods: however, major variations in reliability are to be expected due to differences in the degree of subjectivity among specific measures and the extent to which observers are trained on the fine points of applying specific measures. Peer and SME appraisals based upon substantiating data should be more reliable than those not based on substantiating data.

It is possible, and perhaps it is likely, that subjective ratings of unit performance will have higher reliability scores than direct observations. For example, if a trainer asks ten SMEs to decide if a unit should receive a "go" or a "no go" on the performance of a collective task, these SMEs might agree that performance rates a "no go." On the other hand, questioning these SMEs on the specific aspects of task performance on which the rating was based may find that each rater was responding to a different aspect of performance.

Validity

An important consideration under the topic of validity is that the TTPs for the modernized force are still under development. What we need are product-oriented measures to help validate the more process-oriented measures. An example of how product-oriented aids can be used to validate processes is provided in Figure 6. This figure shows how a decrease in use of smoke is associated with an increase in casualties sustained by the breaching force. In this case the process-oriented measure being validated is not "whether smoke is used." Instead, the process-oriented measure being validated is whether "smoke is employed throughout the breaching operation."

Any of the five types of measurement can be validated against product-oriented measures. The product-oriented measures likely to be of the greatest value in the validation process are direct observations, objective scoring, and peer appraisals.

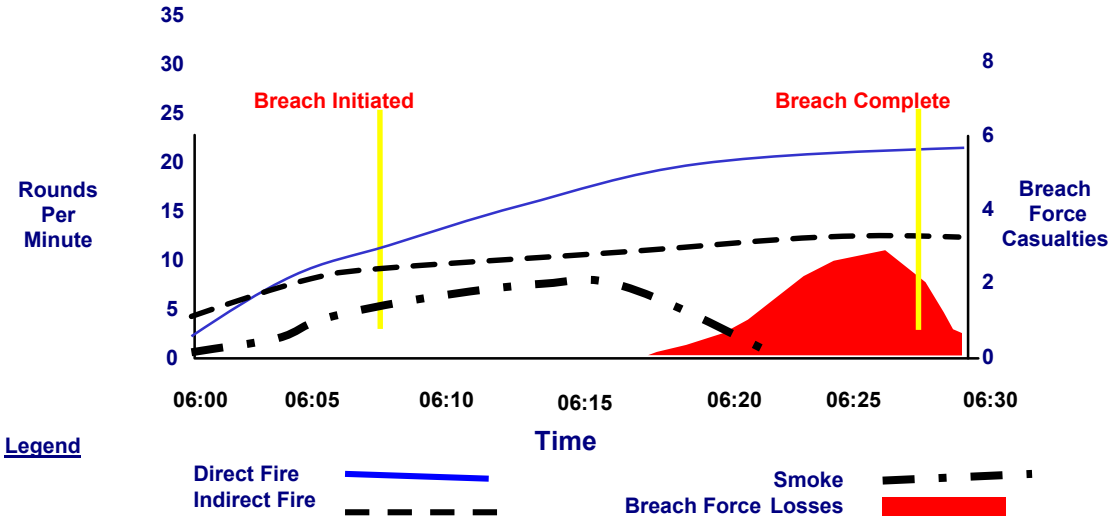


Figure 6. Relationship between use of smoke and protection of breach force.

Documentation

Before units can be expected to try to correct performance problems, the existence of the problems must often be proven to the unit. Direct observation of behavior and objective scoring provide greater opportunities to document specific unit performance strengths and weaknesses. Peer and SME appraisals with substantiating data provide significant opportunities to document strengths and weaknesses, but substantiating data may often be lacking.

Specific Corrective Actions

Concrete, rather than abstract, descriptions of unit strengths and weaknesses are needed to identify corrective actions (Downs, Johnson, and Fallesen, 1987). In general, direct behavioral observations point towards specific, concrete corrective actions. Behavioral observations may identify specific processes that were not employed by a unit, or they may identify a more product-oriented problem (e.g., the unit failed to use supporting fires until after it had already sustained substantial losses during direct fire engagements) and leave it to the unit to identify the cause(s) of the problem.

A drawback in the use of measures based on direct observations is that laundry lists of problems may be produced (e.g., here is a problem in the way the order was issued, here is a problem in preparing fighting positions). This type of output does not help to identify exercises offering a high payoff in terms of training needs addressed. Objective scoring, on the other hand, offers the potential for defining training strategies. For example, objective scoring may show that a unit failed to meet a majority of the criteria for reacting to contact, suggesting it might be beneficial to train on this activity.

All types of appraisals may provide results that are too abstract to meet performance measurement objectives without collecting additional data. For example, a unit leader may conclude that communication is a problem without any clarification (Up the chain of command? Down the chain of command? With specific higher, supporting and adjacent units? Too many/few communications? Critical details not being addressed?). Addressing this problem requires either imposing structure on the appraisal process before the fact (providing examples of behavioral observations that warrant specific ratings and training appraisers how to use this guidance), or collecting data to clarify the appraisal after the fact. Either approach adds greatly to the data collection workload.

Workload

The work required to collect and analyze unit performance measurement data is a major concern of the U.S. Army Training and Doctrine Command (TRADOC). ARI has performed five studies for the Army Training Modernization Directorate (ATMD) that address the issue of observer/ controller (OC) and analyst workloads in the live force-on-force training situation. One of the latest of these studies concerns the use of centralized analysis facilities to support training feedback at multiple sites concurrently (Anderson, Begley II, Arntz, and Meliza, in preparation). Current plans call for applying automation to reduce the workloads associated with preparing training feedback in the live combat training center and home station training environments. Further, the Army is currently automating the AAR aid preparation process in the Close Combat Tactical Trainer (CCTT) environment. Most of these automation activities concern the application of automation to performance measures involving direct behavioral observation, but, as mentioned numerous times in this paper, performance measurement will also require the use of peer, self and SME appraisals.

Sole Source

Many cognitive activities cannot be directly observed, making it necessary to employ self appraisals to find out what happened. The perceived tactical situation and the logic behind a leader's plan for responding to this situation are examples of such cognitive activities. To some extent this limitation will be reduced through the availability of digital situational awareness data and automated decision aids.

Unit SOPs are important variables in controlling and explaining unit activities. In many cases these SOPs are known to the members of a unit but not to outside observers. Again, self appraisals and friendly peer appraisals are important sources of information regarding unit SOPs and their impact on unit performance.

Impacts of Force Modernization

Force modernization will itself influence the process of measuring unit performance. First, the TTPs to be employed by units are still under development, making it likely that certain process-oriented measures of performance will change. Second, use of digital systems, combined with advances in training technology, will make it possible to provide units with real

time feedback regarding unit performance employing measures based upon direct observations. Units will be able to respond to this feedback during the mission planning, preparation and execution process. Third, improved situational awareness will enhance the value of self appraisals. Fourth, in the absence of interventions, the workload required to employ measures based upon behavioral observations will increase substantially (Brown et al., 1998). The increase in workload is due, in part, to the fact that trainers must observe the operators of digital systems, the users of digital system information, interactions among system operators and users, and interactions among operators.

Summary

Direct observations and objective scoring are the most useful types of measures due to their greater reliability and due to the fact that they can document performance problems and point towards specific corrective actions. Direct observations and objective scoring also play a key role in validating other types of measures.

Peer, self, and SME appraisals help to address gaps in the information available from direct observations and objective scoring. To realize the potential value of appraisals, a data collection team must expend significant effort to obtain data that substantiates appraisals.

References

Anderson, L., Begley II, I.J., Arntz, S. & Meliza, L.L. (in preparation). Training Analysis and Feedback Center of Excellence (TAAF-X) (ARI Study Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Brown, B.R., Nordyke, J. W., Gerlock, D. L., Begley, I. J. II, & Meliza, L. L. (1998). Training analysis and feedback aids (TAAF Aids) study for live training support (ARI Study Report 98-04). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A351 107)

Downs, C.W., Johnson, K.M. & Fallesen, J. J. (1987). Analysis of feedback in after action reviews (ARI Technical Report 745) Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Dwyer, D.J., Fowlkes, J.E., Oser, R.L., Salas, E., & Lane, N.E. (1997). Team performance measurement in distributed environments: The TARGETS methodology. IN M.T. Brannick, E. Salas, and C. Prince (Eds.) Team performance assessment and measurement (pp. 137-153). Malwah, NJ: Lawrence Earlbaum Associates.

Gerlock, D. & Meliza, L. "Supporting exercise control and feedback in the digital domain for virtual simulations". Paper presented at the Interservice/Industry Training, Simulation and Education Conference, Orlando, FL, December, 1999.

Mirabella, A, Siebold, G.L., & Love, J.F. (1998, June). Assessment of command performance and feedback in multiforce training. Paper presented at the NATO RTO Workshop, The Human in Command, Kingston, Ontario, Canada.

Morrison, J.E. & Meliza, L.L. (1999). Foundations of the After Action Review Process. (ARI Special Report 42). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

APPENDIX A: Agenda

Workshop on Assessing and Measuring Training Performance Effectiveness

6 SEP

8:00 - 8:30 Check in (coffee/breakfast pastries)

8:30 – 8:45 Introductions: Dr. Jack Hiller, Chief Scientist, TRW, Training and Simulation, Workshop Organizer

8:30- 8:45 Welcome: Robert Seger, ADCS-T TRADOC

8:45- 9:00 Workshop Goals: Dr. Ed Johnson, Director ARI

Presentations

9:00 – 9:15 "The requirement for measuring and assessing training effectiveness"
Dr. Bob Bauer, Dep. Dir., Directorate of Training and Doctrine Development, Armor School & Center

9:15 – 10:00 "Assessment Applications and Advances"
Dr. Eva Baker, Director of the UCLA National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

10:00 – 10:15 Break

10:15 – 11:00 "New Leadership Toolkit"
Stephen Zaccaro, George Mason University

11:00 – 12:00 "Assessing Staff Operations and Functions in Digitized Units"
MG(R) Lon (Bert) Maggart, Research Triangle Institute

12:00 – 1:00 Lunch

1:00 – 5:00 Break into 5 Panel groups to receive presentations

5:00 – 7:00 Break for dinner

7:00 Dinner

Speaker: " Perspectives", LTG(R) Paul Funk, VP Gen. Dynamics Land Systems

Panel 1. "Proficiency Measurement in Technical Training Evaluation"
Co-chairs: Dr. Millie Abell, Chief of Technologies Branch, Futures Training Div. ADCS-T, and Dr. Scott Graham, ARI
Facilitator: Dr. Jerry Childs, TRW
Topics:

- Methods for evaluating on-the-job performance, strengths and weaknesses, Dr. Paul Rossmeissl, AIR
- Modeling and Measuring Situational Awareness, Dr. Scott Graham, Chief ARI Infantry Forces Research Unit and Dr. Michael Matthews, USMA, West Point
- Measuring Performance in Distance Learning, Dr. Robert Wisher, ARI
- Panel discussion on issues, solutions, and possible R&D needs

Panel 2. "Leadership Training and Education"
Co-chairs: COL Chris Sargent, Dir. CAL, Dr. Mike Drillings, Chief ARI RACO
Facilitator: LTG (R) Don Holder
Topics:

- Command Preparation, Opportunities and Needs, TBA
- Training Adaptive Leaders - Are We Ready?, Dr. Karol Ross, HRED and Dr. Jim Lussier, ARI
- Panel discussions

Panel 3. "Staff Training Assessment."
Co-chairs: COL Marven Nickels, Dir. CA&Staff Service Sch., and Dr Kathy Quinkert, Chief ARI TRADOC Scientific Coordination Office
Facilitator: BG(R) William Mullen III, TRW
Topics:

- SIMITAR Assessment, John Metzko and John Morrison, IDA 45 min
- Panel discussions

Panel 4. "Unit Collective Training"
Co-chairs: COL Kent Ervin Dir. Collective training Directorate, CAC, Dr.Goldberg, ARI
Facilitator: COL (R) John Johnston, TRW
Topics:

- Unit Training Measurement and Evaluation Issues, Henry Simpson, OSD
- Making the case for training systems (CCTT) evaluation, Dr. Goldberg, Chief ARI Orlando
- Panel discussions

Panel 5. "Performance Measurement and Assessment Issues"
Co-chairs: MAJ Steve Ellison, TRADOC DCS-T TDAD, & Dr. Elizabeth Brady, ARI
Facilitator: Dr. Ward Keesling, PRC
Topics:

- Formal tests and measures, including issues of reliability and validity, Andy Rose, Chief Scientist, Amer. Inst. For Research 45 min
- Strengths and Weaknesses of Alternative Measures: Rating by Direct Observation, Objective Scoring of Results, Self-appraisal, Peer Appraisal, & SME Judgment, Dr. Meliza, ARI
- Fixing Inter-rater Reliability, Dr. Frank Apicella, Technical Dir., AEC
- Panel discussions

7 SEP

8:00 – 8:30 Coffee & pastries

8:30 - 12:00 Reform into Panel groups

Panels will each finish discussions and then prepare reports to main body.
Facilitators will work to gain coverage of consensus for technical strengths and weaknesses of current methods, identify issues, tentative solutions and recommendations, and significant minority views.

12:00 – 1:00 Lunch

1:00 Begin Panel reports (30 minutes each)

3:00 – 3:15 Break

3:15 – 3:45 Final Panel report

3:45 – 4:00 Workshop Summary

Appendix B

Attendee	Organization
Abbey, SFC Michael	ATEC
Abell, Dr. Millie	Futures Trng Div, TRADOC
Anderson, Edwin P.	USASSI Dir. Of Training
Apicella, Frank	Army Evaluation Command
Askey, Ronald	DOTD, USAJFKSWCS
Baker, Dr. Eva	UCLA National Center for Research
Bauer, Dr. Bob	DTDD, US Army Armor Center
Beasley, LTC Dan	Dpt of Tng, Plans & Evaluations, Ft Eustis
Berg, Mary	TRADOC Eval & QA Program Mgr
Billups, Deborah	ATSC, OLPD
Brady, Dr. Elizabeth	ARI
Campbell, Rebecca	DOTD, USAJFKSWCS
Carberry, Ed	PM ASAT, ATISD, USATSC
Childs, Dr. Jerry	TRW - ABQ
Coose, Phil	TRW - ABQ
Crandell, James	Warrior T, Ft. Hood, TX
Dawson, Brenda B.	Development Division, Army Transportation School
Drillings, Dr. Mike	ARI, Office Rsch & Adv Concepts
Ellison, MAJ Steve	TRADOC Eval & QA Program Mgr
Ervin, COL Kent	Collective Training Directorate, CAC
Faber, Terry D.	(ATIC-ATM) ATSC, Ft. Eustis, VA
Feldmeier, Howard	IBCT
Ferris, George	School of Americas, Ft Benning, GA
Fuglestadt, Tom	IBCT
Funk, LTG (R) Paul	VP Gen. Dynamics Land Systems
Goldberg, Dr. Steve	ARI, STRICOM
Graham, Dr. Scott	ARI, Ft. Benning, GA
Hamilton, CPT Andy	ABCS Team ATSC, WarMod Div.
Hardy, Carlton	ABCS Team ATSC, WarMod Div.
Hiller, Dr. Jack	TRW - Monterey, CA
Holder, LTG (R) Don	TRW - Killeen, TX
Holtz, Louis W.	Directorate of Training, USAES
Hunter, Brian	IBCT
Johnson, Dr. Edgar	Director, ARI
Johnston, COL (R) John	TRW - Killeen, TX
Keenan, Leo	AB TECH, STRICOM
Keesling, Dr. Ward	PRC - Monterey, CA
Larsen, James E.	DCST Recruiting Directorate, TRADOC
Lesjak, Neta T.	AMEDDCS&S, Ft. Sam Houston,TX
Livingston, Elaine	Infantry School
Luker, Mark	Lead Evaluator for IBCT
Lussier, Dr. Jim	ART - Ft. Knox Team Leader
MacAllister, Mac	Eval & QA Program Mgr, TRADOC
Maggart, MG (R) Lon (Bert)	Research Triangle Inst.

Meliza, Dr. Larry	ARI, Orlando, FL (STRICOM)
Melton, Bill	TRACOC TDAD
Metzko, John	SIMITAR ASSESSMENT
Mitchell, Mary E.	US Maneuver Support Center Warrior Dept
Morrison, John	SIMITAR ASSESSMENT
Mullen, BG (R) Bill	TRW - Monterey, CA

Attendee	**Organization**
Nicholson, Nigel	Army Eval Ctr (ATEC)
Nickels, COL Marven	CAS3, Ft Leavenwoth, KS
Nollette, COL John A.	TASS Integration Element, Ft Sill, OK
Olson, COL Chris	TRADOC
Parker, MAJ Walt	Schools Division, ITD, DCST, TRADOC
Parodi, CPT Mike	ABCS Team ATSC, WarMod Div.
Pettie, Alan	ATSC, OLPD Trng Spt Ctr
Pittman, James	Redstone Arsenal
Powell, Alyce F.	DOTD WarMod
Quinkert, Dr. Kathy	ARI TRADOC
Rauchfuss, SFC Gary	School of Music
Riedel, Sharon	ARI - Research Psychologist
Ronneberg, Ron	Director War Mod Div.
Rose, Dr. Andy	American Institute for Research
Ross, Dr. Karol	ARI - Ft Sill, OK
Rossmeissl, Dr. Paul	American Institute for Research
Schoch, Bruce	HQ CASCOM
Seko, Dave	Chief, ABCS Team
Serio, Rachel L.	TRADOC Eval & QA Program Mgr
Shrode, Tricia	Armor School
Simpson, Dr. Henry	OSD DMDC
Smith, Sharon	PEC
Snel, LT Joseph	Army Reserve Readiness Trng Ctr
Snider, Floyd M.	Aviation Center, DOTDS, Ft. Rucker
Snyder, Mary R.	Army Reserve Readiness Trng Ctr
Taylor, Don	IBCT
Tierney, Diana	Futures Trng Div, TRADOC
Townsley, Norma J.	HQ CASCOM
Tyler, Edward C.	TRADOC
Van Deren, Richard W.	Development Division, Army Transportation School
Wagner, Hilde	ATSC, OLPD Trng Spt Ctr
Walton, Barbara H.	ATZH-DTS Ft. Gordon, GA
Wampler, Rich	TRW - Columbus, GA
Wardell, Connie	Armor School
Wauthier, Jerome	MP School, Ft. Leonard Wood, MO
White, Bob	TDAA, DCST, TRADOC
Wightman, Dennis	ARI - Rotary Wing Aviation Rsch Unit
Williams, Vivian	US Army Logistics Management College Ft. Lee, VA
Wisher, Dr. Robert	ARI
Wright, John	ATMD (TRADOC)

Zaccaro, Steve	Professor, George Mason University
Zinn, Dr. Will	ATSQ-LAC-P, Ft Eustis, VA

APPENDIX C: Panel Summaries

APPENDIX D: Workshop Presentations

Keynote Presentations

Panel 1: Proficiency Measurement in Technical Training Evaluation.

Panel 2: Leadership Training and Education.

Panel 3: Staff Training Assessment.

Panel 4: Unit Collective Training.

Panel 5: Performance Measurement and Assessment Issues.